PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :
C12Q 1/68

A2

(11) International Publication Number: WO 99/11823

(43) International Publication Date: 11 March 1999 (11.03.99)

(21) International Application Number: PCT/US98/18392

(22) International Filing Date: 4 September 1998 (04.09.98)

(30) Priority Data:
08/925,816 5 September 1997 (05.09.97) US

(71) Applicant (*for all designated States except US*): SIDNEY KIMMEL CANCER CENTER [US/US]; Suite 200, 10835 Altman Row, San Diego, CA 92121 (US).

(72) Inventors; and
(75) Inventors/Applicants (*for US only*): McCLELLAND, Michael [US/US]; 804 Avenida de San Clemente, Encinitas, CA 92024 (US). PESOLE, Graziano [IT/IT]; Via Fanelli, 206/L, I-70126 Bari (IT).

(74) Agents: WONG, James, J. et al.; Campbell & Flores LLP, Suite 700, 4370 La Jolla Village Drive, San Diego, CA 92122 (US).

(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, GH, GM, HR, HU, ID, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
Without international search report and to be republished upon receipt of that report.

(54) Title: SELECTION OF PCR PRIMER PAIRS TO AMPLIFY A GROUP OF NUCLEOTIDE SEQUENCES

(57) Abstract

The present invention provides a method of determining a set of primer pairs for amplifying a group of related nucleotide sequences. A method of the invention is performed by identifying a group of related nucleotide sequences; generating the set of primers that matches each of the related nucleotide sequences; determining for each systematic pairing of each primer which of the related nucleotide sequences are amplified; and selecting from the systematic pairings a subset which amplifies all of the related nucleotide sequences. The invention also provides a method of using a set of primer pairs, which amplify a group of related nucleotide sequences, to identify nucleotide sequences related to the original group of nucleotide sequences. The invention also provides a computer apparatus for carrying out the computer-executed steps of the method. The invention further provides a computer program product comprising a signal bearing media for carrying out the method.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

SELECTION OF PCR PRIMER PAIRS TO AMPLIFY A GROUP OF NUCLEOTIDE SEQUENCES

This invention was made with government support under grant numbers CA68822, NS33377 and AI34829 awarded by the NIH. The government has certain rights in the invention.

BACKGROUND OF THE INVENTION

FIELD OF THE INVENTION

The present invention relates generally to methods of amplifying nucleotide sequences and more specifically to methods of identifying sets of primer pairs sufficient to accomplish the amplification.

BACKGROUND INFORMATION

Every living organism requires genetic material, deoxyribonucleic acid (DNA), to pass a unique collection of characteristics to its offspring. DNA is made of two strands of nucleotide building blocks; the two strands bind, or hybridize, much like a zipper and form a double helix. Genes are discreet segments of the DNA and provide the information required to generate a new organism. Even simple organisms, such as bacteria, contain thousands of genes, and the number is many fold greater in complex organisms such as humans. Understanding the complexities of the development and functioning of living organisms requires knowledge of these genes. However, the amount of DNA that can be isolated for study has often been limiting.

A major breakthrough in the study of genes was the development of the polymerase chain reaction (PCR). PCR "amplifies" genes or portions of genes by making many identical copies, allowing isolation of genes from very tiny amounts of DNA. PCR requires the design of primers composed of short stretches of nucleotides that bind, or hybridize, to discreet segments of the gene. Using a DNA polymerase enzyme, the gene or portion of the gene that is adjacent to the bound primers is then copied many times, generating large quantities of material that can be used for further studies, such as identifying genes expressed abnormally in cancer cells.

The design of PCR primers is relatively straightforward when the sequence of the gene of interest is known or when the number of sequences to be amplified is small. However, particular circumstances can make the design of PCR primers a difficult task. For example, it would be advantageous to be able to identify new related genes where the sequence of some related genes are already known. However, design of PCR primers requires that the sequence to be amplified is known. The length of primers is also critical for successful PCR amplification. For example, primers of sufficient length will selectively amplify a known gene but will likely be too specific to amplify related family members. On the other hand, if primers are too short, they will amplify many unrelated genes and generate a high background. In addition, PCR amplification of a large number of genes requires the design of a large number of primers. Thus, a need exists for a method to select a set of primers that can amplify all of the genes of interest. The present invention satisfies this need and provides additional advantages as well.

SUMMARY OF THE INVENTION

The present invention provides subsets of primers sufficient to amplify a group of related nucleotide sequences. The subset of primers can be less than the maximum number of two primers per nucleotide sequence required to amplify a group of related nucleotide sequences. For example, the invention provides a subset of primers where the number of primers is less than or equal to the number of related nucleotide sequences in the group. The primers in the subset can be limited to a specific length or range of G+C content, or the subset of primers can exclude primers that amplify an undesirable nucleotide sequence.

The present invention also provides a method of determining a set of primer pairs for amplifying a group of related nucleotide sequences. For example, the invention provides a method of determining a set of primer pairs for amplifying a group of structurally related nucleotide sequences that encode members of the human nuclear receptor family. A method of the invention is performed by identifying a group of related nucleotide sequences; generating a set of primers that match each of the related nucleotide sequences; determining for each systematic pairing of each primer which of the related nucleotide sequences are amplified; and selecting from the systematic pairings a subset of primers sufficient to amplify all of the related nucleotide sequences.

The invention additionally provides a method of identifying an amplified nucleotide sequence that is related to an original group of related nucleotide sequences. For example, a set of primers that samples a known group of nucleotide sequences that is induced by

TGF-β can be used to identify previously unknown related nucleotide sequences induced by this agent.

The invention further provides a computer apparatus comprising a processor, main memory in communication with the processor, and a primer pair selector in communication with main memory for carrying out the computer-executed steps of identifying a group of related nucleotide sequences; generating a set of primers that match each of the related nucleotide sequences; determining for each systematic pairing of each primer which of the related nucleotide sequences are amplified; and selecting from the systematic pairings a subset of primers which is sufficient to amplify all of the group of related nucleotide sequences.

The invention also provides a computer program product for determining a set of primer pairs sufficient to amplify a group of related nucleotide sequences comprising means for identifying a group of related nucleotide sequences; means for generating a set of primers that match each of the related nucleotide sequences; means for determining for each systematic pairing of each primer which of the related nucleotide sequences are amplified; means for selecting from the systematic pairings a subset of primers which can amplify all of the related nucleotide sequences; and signal-bearing media containing the means for the identifying, generating, determining and selecting.

The present invention is best carried out as software operating in a computer, but one skilled in the art will recognize that it can be carried out as hardware or as a combination of hardware and software.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 shows a block diagram of the computer system of the preferred embodiment.

Figure 2 shows the flowchart that describes the operation of the present invention.

Figures 3 and 4 show the flowcharts that depict the computer-executed steps for carrying out the method of the invention.

Figure 5 shows a program product for performing the method of the invention.

DETAILED DESCRIPTION OF THE INVENTION

The present invention provides subsets of primers sufficient to amplify a group of related nucleotide sequences. The subset of primers can be less than the maximum number of two primers per nucleotide sequence required to amplify a group of related nucleotide sequences. For example, the invention provides a subset of primers where the number of primers is less than or equal to the number of related nucleotide sequences in the group. The primers in the subset can be limited to a specific length or range of G+C content, or the subset of primers can exclude primers that amplify an undesirable nucleotide sequence.

The present invention also provides a method of determining a set of primer pairs for amplifying a group of related nucleotide sequences. For example, the invention provides a method of determining a set of primer pairs for amplifying a group of structurally

related nucleotide sequences which encode members of the human nuclear receptor family. A method of the invention is performed by identifying a group of related nucleotide sequences; generating a set of primers that match each of the related nucleotide sequences; determining for each systematic pairing of each primer which of the related nucleotide sequences are amplified, for example, by generating a matrix that ranks the primers and nucleotide sequences by the number of matches to the other; and selecting from the systematic pairings a subset of primers that is sufficient to amplify all of the related nucleotide sequences.

An advantage of the invention is that a set of primers is identified that allows for amplification of all or most of a group of related nucleotide sequences, while minimizing the number of primer pairs required for amplifying the group. As used herein, the term "primer" refers to an oligonucleotide sequence of any size that can be used to amplify a nucleotide sequence. In a preferred embodiment, the primers are about 5 to about 50 nucleotides in length, generally about 8 to about 12 nucleotides in length. As used herein, the term "group of related nucleotide sequences" refers to DNA or RNA molecules that share a common feature. The common feature is a feature of interest to an investigator, for example, a group of related nucleotide sequences that share one or more common structural features, such as the DNA binding domains shared by members of the human nuclear receptor family. A group of related nucleotide sequences also can share a common feature such as being involved in DNA repair or apoptosis or being induced in response to a stimulus such as TGF-β. If desired, a shared common feature of a group of related nucleotide sequences can define a large group of nucleotide

sequences such as human mRNA sequences, which share the common feature of being expressed in a particular human cell. As set forth herein, a group of related nucleotide sequences can be compiled as a list (see, for example, Table II, Table V, Table VII and Table IX). Similarly, a set of primers can be compiled as a list.

A common feature of a group of related nucleotide sequences also can be a common function. For example, a group of related nucleotide sequences can encode proteins that share a common function, which can be any biological function. Such proteins include receptors, for example cell surface receptors, cytoplasmic receptors such as steroid hormone receptors, and nuclear receptors; secreted proteins; non-secreted proteins; hormones such as peptide hormones; and signal transduction proteins. In addition, a group of related nucleotide sequences can be genes regulated during certain conditions, such as genes regulated during cell growth or regulating cell growth; as well as genes regulated during development, during pathogenesis, by pathogens, by drugs, by stress, by radiation exposure, during tissue repair, during senescence or during aging. Also, tumor suppressor genes, DNA repair genes, DNA replication genes, or DNA repair and replication genes can be groups of related nucleotide sequences. Additionally, genes responsible for resistance to therapy, resistance to drugs or for increased sensitivity to therapy or drugs also can be a group of related nucleotide sequences. Such groups of related nucleotide sequences as these herein described can be broadly applicable to plant biology, agriculture, medicine and reproduction, veterinary medicine, microbiology and environmental sciences.

Primers identified by a method of the invention match with specific regions of nucleotide sequences. As used herein, the term "match" refers to a primer that is 100% identical to a region of a nucleotide sequence. The primers identified herein can themselves function, for example, as PCR primers in a test tube or reaction vessel, or can be part of a longer oligonucleotide sequence. Thus, a primer selected using the disclosed method can comprise a longer oligonucleotide that, in fact, is used to perform PCR. For example, the subset of primers that amplify the human nuclear receptor family shown in Table III can be synthesized and used to amplify nucleotide sequences related to this family. To actually perform PCR on a cDNA sample, however, additional oligonucleotide bases can be added to the an end of the primers shown in Table III. These additional bases, which can be arbitrary bases, aid in stabilizing the hybridization of PCR primers to the cDNA sequence. Thus, the subset of primers identified using a method of the invention provide a core sequence, identical to a region of a DNA or RNA molecule, that allows amplification of such a nucleotide sequence, even though additional oligonucleotide bases can be added to the ends of these identified primers to facilitate performing PCR in a test tube.

One method of amplifying nucleotide sequences involves PCR. However, the primers disclosed herein can be used to amplify nucleotide sequences by any method that amplifies nucleotide sequences. Examples of such amplification methods include the ligase chain reaction (LCR), self-sustained sequence replication (3SR), beta replicase reaction, for example, Q-beta replicase reaction, phage terminal binding protein reaction, strand displacement amplification (SDA) or nucleic acid

sequence-based amplification (NASBA) also can be used to amplify nucleotide sequences using the primers of the invention (Trippler et al., J. Viral. Hepat. 3:267 (1996); Hofler et al., Lab. Invest. 73:577 (1995); Tyagi et al., Proc. Natl. Acad. Sci. USA 93:5395 (1996); Blanco et al., Proc. Natl. Acad. Sci. USA 91:12198 (1994); Spears et al., Anal. Biochem. 247:130 (1997); Spargo et al., Mol. Cell. Probes 10:247 (1996); Gobbers et al., J. Virol. Methods 66:293 (1997); Uyttendaele et al., Int. J. Food Microbiol. 37:13 (1997); and Leone et al., J. Virol. Methods 66:19 (1997)), each of which is incorporated herein by reference.

The present invention is based on the understanding that a group of related nucleotide sequences share common stretches of nucleic acid sequence. For example, in the simplest case, nucleotide sequences encoding structurally related proteins share regions of homology that encode conserved domains. Thus, the related nucleotide sequences in such a group are all structurally related. However, as disclosed herein, a group of related nucleotide sequences need not encode proteins that are structurally similar, but also can be nucleotide sequences that are commonly induced in response to a stimulus, for example, DNA damage, or exposure of a cell or organism to a drug or other chemical agent. In this case, some or all of the related nucleotide sequences in a group are not structurally related.

The present invention recognizes that there is a statistical probability that a group of related nucleotide sequences share one or a few common short stretches of nucleotide sequence. For example, a primer eight nucleotides in length would be expected to occur,

statistically, about once every 65,000 base pairs (4^8). Using a method as disclosed herein, a sequence corresponding to an 8-mer primer occurred 34 times in the identified group of 44 human nuclear receptors (Table II), whereas statistics would have predicted that the 8-mer would occur only once in this list of related nucleotide sequences. This result demonstrates that a group of related nucleotide sequences share common short stretches of nucleotide sequence, which a method of the invention identifies.

In one embodiment, the invention provides a means to identify a set of primers that samples a group of related nucleotide sequences. As used herein, the term "samples" refers to the ability of a primer pair to hybridize to and amplify a nucleotide sequence by PCR. The set of primers can be a minimal or near minimal set, the minimal set being a set of primers containing the fewest number of primers sufficient to amplify a group of related nucleotide sequences. The set of primers is selected from all possible primers of a given size or range of sizes. For example, 8-mers would generate 65,536 (4^8) possible primers, whereas 10-mers would generate 1,048,576 (4^{10})possible primers.

The disclosed method provides a means for identifying a group of related nucleotide sequences; generating a set of primers that matches each of the group of related nucleotide sequences, for example by selecting a primer size such as an 8-mer and determining for each primer which of the related nucleotide sequences are amplified; determining for each systematic pairing of each primer, which of the related nucleotide sequences are amplified, for example, by generating a matrix that ranks the primer pairs and nucleotide sequences by the

number of matches to the other; and selecting from the systematic pairing a subset of primers sufficient to amplify all or most of the group of related nucleotide sequences. As used herein, the term "systematic pairing" refers to pairing any given primer in a set with all primers in the set, including itself. For example, the invention provides a method to amplify genes in the human nuclear receptor family (see Table II).

The invention provides a means to identify a subset of primers that amplify a group of related nucleotide sequences. Visual inspection of a group of related nucleotide sequences allows identification of a set of primers sufficient to amplify some groups of related nucleotide sequences. However, visual inspection for identification of PCR primers that amplify a group of related nucleotide sequences is efficient only when the number of related nucleotide sequences is small, for example, 5 nucleotide sequences or fewer, and the nucleotide sequences are relatively short. When a group of related nucleotide sequences contains a large number of nucleotide sequences or if the nucleotide sequences are long, however, a computer process conveniently performs the identification of the set of primers. As used herein, the term "computer process" refers to a method for carrying out computer-executed steps. For example, if the group of nucleotide sequences contains at least 10 nucleotide sequences or more, or 25 nucleotide sequences or more, particularly 50 nucleotide sequences or more, a computer process provides an efficient method to identify primers that amplify the group of related nucleotide sequences.

A computer process can compile a list of all possible primers composed of short nucleotide sequences

that can function as PCR primers. The primers range in size from about 5 to about 50 nucleotides, for example, about 8 to about 12 bases long, and are generally about 8 or 9 bases in length. The computer process provides an advantage in that a large number of possible primers can be considered, for example, the 65,536 (4^8) possible 8-mers, and the primers can be restricted by specific criteria, such as restricted to a predetermined range of G+C content or restricted to exclude matches to undesirable nucleotide sequences, for example, ribosomal RNA sequences. As exemplified herein, a computer process, using the program WORDUP, was used to compile a list of all possible primers of various lengths (see Example I; Pesole et al., Nucleic Acids Res. 20:2871 (1992)), which is incorporated herein by reference.

Primers as short as 5 bases have been used for PCR (Caetano Anolles et al., Biotechnology 9:553 (1991)). For example, a large group of related nucleotide sequences can have common 5-mer primers that amplify all or most of the related nucleotide sequences. However, while very short primer sequences will occur frequently in related nucleotide sequences, they also statistically are likely to occur frequently in other unrelated nucleotide sequences. Therefore, very short primer sequences do not always provide the desired selectivity for amplifying the group of related nucleotide sequences. The primer length is also constrained by the statistical frequency of any given nucleotide sequence. For example, a 10-mer would occur about once every 1,000,000 base pairs (4^{10}). In the list of 44 genes of the human nuclear receptor family (Table II), the most common 10-mers occurred only a few times and were confined to regions of conserved domains. Thus, for determining a subset of primers sufficient to amplify the human nuclear receptor

family, the primers were limited to lengths no greater than 9-mers.

If desired, primers that match undesirable nucleotide sequences can be eliminated from the compiled list of possible primers. Abundant nucleotide sequences such as mitochondrial DNA and ribosomal RNA or dispersed repetitive elements can lead to high background if primers in the set hybridize to these sequences during PCR. Thus, primers that match such abundant sequences can be eliminated from the compiled list of possible primers.

In humans, the most abundant dispersed repetitive sequences are Alu and LINE elements (Quentin, Genetica. 93:203 (1994); Kariya et al., Gene 53:1 (1987); Amariglio and Rechavi, Environ. Mol. Mutagen. 21:212 (1993)). The ribosomal RNA sequences and mitochondrial RNA sequences, transcribed from mitochondrial DNA, constitute a substantial and variable proportion of the RNA population of a cell, even after poly(A) RNA selection. A list of human ribosomal RNA sequences, human mitochondrial DNA sequences, and eight representative Alu elements were compiled (see Table I). A list of 99 mRNA sequences that carry fragments of LINE elements was also compiled (Hattori et al., Nature 321:625 (1986)),(see Table I). LINE elements generally are found in a 5′ truncated form of mRNA sequences and, where LINE elements occur, they generally are found in the 3′ untranslated region of mRNA sequences. Primers that match such sequences can be eliminated from the compiled list of possible primers. For example, using the human nuclear receptor sequences shown in Table II, all primers that matched with either strand of the human nuclear receptor sequences were compared to the ribosomal RNA, mitochondrial DNA and Alu sequences listed in

Table I and those primers that matched with either strand of these abundant nucleotide sequences were removed from the compiled list of possible primers. In addition, any primer that occurred three or more times in the list of LINE elements in Table I was removed from the compiled list of possible primers. After subtraction of abundant nucleotide sequences, only primers greater than 7-mers remained in the compiled list of possible primers. Because the vast majority of the 16,384 (4^7) possible 7-mer primers occurred in the sequences listed in Table I, the primers were limited to lengths no shorter than 8-mers for determining a subset of primer pairs sufficient to amplify the human nuclear receptor gene family, the human G-protein coupled receptor gene family, human apoptosis-associated genes and human DNA repair and replication genes.

For the groups of related nucleotide sequences containing 44 to 113 nucleotide sequences described in Examples II through V, 8-mer primers gave slightly better sampling of the related nucleotide sequences than 9-mer primers after removing primers that occur in abundant RNAs. However, for very large groups of related nucleotide sequences, 9-mers, or longer oligos, have a better chance of occurring in a sufficient number of related nucleotide sequences in a group.

Other undesirable nucleotide sequences can include abundant mRNA sequences. For example, primers that match known abundant mRNA sequences that are constitutively expressed can be subtracted from the compiled list of possible primers. High background of specific sized PCR products of the abundant mRNA sequences can obscure detection of similar sized PCR products of less abundant mRNA sequences. Removal of

background contributed by the abundant mRNA sequences, whose PCR products would have a similar size compared to the less abundant mRNA sequences, would increase the likelihood of detecting amplified products from the less abundant mRNA sequences.

Accumulating information on the relative rank of abundance of mRNAs in cells can allow the identification of the top 100 or 500 mRNAs that occur in most cell types in an organism, for example, humans. A list of such genes can be used to exclude primer pairs from consideration if, for example, more than one of these more abundant RNAs was likely to give a PCR product with a primer pair. To date, most of the publicly available information is confined to genes that are differentially expressed. The level of expression of genes that are not differentially expressed has not been made publicly available. However, information from projects such as the Cancer Genome Anatomy Project and the Human Genome Initiatives EST project can be used to determine expression of genes in an undifferentiated state. Such information can be used to develop a list of the 100 most consistently highly expressed genes, which can then be excluded from a list of primers.

Alternatively, experiments have been routinely performed with total cDNA as a control probe for arrays. Genes that are detected by a total cDNA probe and are not differentially expressed among various cell lines can be used to generate a catalog of consistently highly expressed genes that can be utilized to exclude primers that match abundant mRNAs.

It is possible that very common genes will predominate in a fingerprint even if such common genes do

not perfectly match with the primer pairs used. Therefore, primers that do not perfectly match with a common gene but that could still be used to amplify the common gene can be excluded from the list of primers. In particular, primers mismatched with the 5'- most specified base can be excluded. If such a primer pair samples more than a threshold number of these common genes, it can be excluded.

The exclusion of primers that occur in abundant RNAs does not exclude primers that are common in the other 10,000 or more mRNAs in the cell that are not in the group of related nucleotide sequences of interest. Some of these other sequences will also have perfect or near-perfect matches with the primers, oriented correctly for PCR and at an appropriate distance apart. Thus, primers selected by the method of the invention increases the probability that the related nucleotide sequences of interest are given an opportunity to be sampled but will not exclude other nucleotide sequences from being sampled in the same mixture of PCR products. However, the invention is not intended to select primers that have perfect matches exclusively with a number of the sequences of interest while having no matches in other sequences. Rather, the methods of the invention are directed to ensure that the related nucleotide sequences of interest are among the best matches so as to maximize the chance that these sequences will occur in a mixture of PCR products.

Although it is possible that a primer pair will sample many other mRNAs, the mixture of PCR products amplified by such a primer pair is still enriched for the intended mRNAs, even if other undesired RNAs are also sampled. Consequently, these PCR reaction mixtures can

be effective probes in differential hybridization experiments against clones for the expected mRNAs because the complexity of the probe is much lower than total cDNA. Such PCR reaction mixtures are of particular interest in strategies that array clones or oligonucleotides on chips where the complexity of the probe can be a limiting factor in detecting rare transcripts.

Additional constraints can be imposed on the primers, as desired. For successful PCR, the stability of primer and template interactions should be considered and both primers in a primer pair need to be matched with regard to their melting temperature on the template. Primers that are A+T rich interact less strongly than G+C rich primers. Thus, a window of G+C content of the primers can be imposed, such as requiring a G+C content of about 20 to about 100%, in particular a G+C content of about 50% to about 90%. As used herein, the term "window of G+C content" refers to a range of percent composition for the nucleotides in the primer. Using a window of G+C content provides the advantage that the melting temperature of a pair of primers can be more closely matched to allow both primers to participate in the PCR reaction with similar efficiency.

A computer process can rank the primers in order of the number of related nucleotide sequences to which the primer matches. For example, the computer process generates a set of primers that amplifies all or most of a group of related nucleotide sequences by selecting a specified number of primers that matches the largest number of nucleotide sequences in the group of related nucleotide sequences and creating a set of primers composed of the selected number of primers and

the complement of those primers. The top ranked primers, for example, the 30 primers that match the largest number of nucleotide sequences, thus yields a list containing a set of 60 primers composed of the 30 top ranked primers and their complements. Each primer in the set of primers is paired with each other primer in the set of primers, and all of the theoretical PCR products are determined for each pair of primers and each nucleotide sequence. In the case where a primer pair generates more than one PCR product from a given nucleotide sequence, the shortest PCR product is used, although the computer process records all PCR products generated. A matrix is generated that ranks 1) the related nucleotide sequences by the number of primer pairs that generate a simulated PCR product, and 2) the primer pairs by the number of related nucleotide sequences that are successfully sampled. The matrix is used to generate a subset of primers that can amplify all, or nearly all, of the group of related nucleotide sequences.

The invention entails determining for each systematic pairing of each primer which of the related nucleotide sequences is amplified. A nucleotide sequence in the matrix is selected by predetermined criteria as the first nucleotide sequence submitted to determine a primer pair that samples the nucleotide sequence. As used herein, "submitted" refers to analyzing a particular sequence using, for example, a computer process. The primer pair that generates a simulated PCR product with the first nucleotide sequence is compared to all other related nucleotide sequences in the matrix. If more than one primer pair can generate PCR products from the first selected nucleotide sequence, the primer pair that samples the largest number of related nucleotide sequences is selected. The primer pair that samples the

first selected nucleotide sequence and all of the related nucleotide sequences sampled by this primer pair are removed from the matrix. The next nucleotide sequence is selected, and the primer pair that samples the second selected nucleotide sequence and all other related nucleotide sequences sampled by this primer pair are removed from the matrix. This process is repeated until a subset of primers is selected that would theoretically PCR amplify all of the related nucleotide sequences. For example, a set of primer pairs that amplifies the human nuclear receptor family is shown in Table III.

If desired, additional constraints can be imposed on the primer pairs. The computer process can require that the PCR products be limited to a predetermined size, for example greater than 100 base pairs and less than 1000 base pairs. Furthermore, the computer process can require that primer pairs generate PCR products that differ by a predetermined size range, such as ± 3 base pairs. Such primer pairs can be advantageous, for example, if the PCR products are to be resolved by gel electrophoresis. Thus, PCR products of sizes that are impractical to separate using predetermined analytical techniques will not be generated. Primer pairs also can be limited to those primer pairs that generate a minimum number of different sized PCR products such as three or more different sized PCR products. In addition, the different PCR products can be limited to PCR products derived from different nucleotide sequences and can be required to amplify a minimum number of nucleotide sequences, such as at least three nucleotide sequences. Also, PCR products can be limited to those with different primers at each end.

The analysis of PCR products on a gel is limited by the available resolution of the gel, by the production of artifactual products from internal priming of the products of interest, and by spurious priming from other genes that are not of interest. Some of these limitations of analyzing PCR products is overcome by analysis of nucleic acids on arrays. For example, size of PCR products is not important for resolution on an array. Also, internal priming on a group of related nucleotide or on unrelated nucleotide sequences present in a sample is not problematic when the level of expression is monitored on an array rather than on a gel. The array effectively "purifies" the products of interest by hybridization to the correct clone. As such, primers to be used to probe arrays need not be limited to a given size differential of PCR products. However, the methods of the invention are still advantageous for identifying primers for analyzing PCR products on arrays because the number of primers can be significantly reduced as described below.

The systematic pairing of primers maximizes the number of related nucleotide sequences that can be sampled with a set of primers. For example, the nucleotide sequence recognized by the largest number of primer pairs can be selected as the first nucleotide sequence submitted to determine a primer pair that samples the nucleotide sequence. The primer pair that generates simulated PCR products with the first nucleotide sequence is compared to all other related nucleotide sequences in the matrix. The primer pair that samples the first selected nucleotide sequence and all of the related nucleotide sequences in the matrix sampled by this primer pair are removed from the matrix. The next nucleotide sequence is selected, in this case, the second

selected nucleotide sequence is the nucleotide sequence remaining in the matrix that is sampled by the largest number of primer pairs. The primer pair that samples the second nucleotide sequence and all other related nucleotide sequences sampled by this primer pair are removed from the matrix. This process is repeated until a subset of primer pairs is compiled that theoretically would result in PCR amplification of all, or nearly all, of the related nucleotide sequences.

An advantage of the invention is that a subset of primers can be identified that provides amplification of a maximal number of related nucleotide sequences, while minimizing the number of primer pairs required for the amplification. An additional advantage is that primer pairs in the subset can be limited by specific criteria. For example, a set of primer pairs can be identified that generates more than one PCR product for each nucleotide sequence, thus assuring that at least two PCR products will be amplified for each nucleotide sequence and providing maximal opportunities to identify PCR products specific for a nucleotide sequence.

To maximize the likelihood that all of the related nucleotide sequences will be amplified at least a minimum number of times, a set of primer pairs that amplifies the nucleotide sequences for at least that minimum number of times is generated. For example, any nucleotide sequences that are not amplified, or are amplified only once, by the set of primer pairs generated as outlined above can be identified and compiled into a new list of nucleotide sequences for analysis, which can be by a computer process. A new set of primers is generated that will match this new list of related nucleotide sequences. This new set of primers can be

added to the first set of primers and submitted to determine a primer pair that samples a nucleotide sequence in the original list of related nucleotide sequences to generate a new matrix that amplifies all, or nearly all, of the group of related nucleotide sequences. This process of identifying nucleotide sequences that are not amplified or are amplified only once by the set of primer pairs can be repeated until all of the related nucleotide sequences have been amplified. Any nucleotide sequence for which primer pairs have not been identified by some number of repeats of this process can be removed from the list or a specific primer or primer pair can be designed, for example, by visual inspection.

Another approach to maximizing the likelihood that all of the related nucleotide sequences will be amplified involves selecting the nucleotide sequence recognized by the fewest number of primer pairs as the first nucleotide sequence submitted to determine a primer pair that amplifies the nucleotide sequence. The method of identifying and removing sampled nucleotide sequences from the matrix is used to identify primer pairs that theoretically amplify all or most of the group of related nucleotide sequences. The identified primer pairs define a subset of primers, that can be a minimal or near minimal set, sufficient to amplify a group of related nucleotide sequences. The process of identifying any nucleotide sequences that are not amplified, or are amplified only once, by the set of primer pairs and of generating a new matrix that allows amplification of these related nucleotide sequences can be repeated until all of the related nucleotide sequences have been amplified, or the process can be terminated at an appropriate point. Using such an approach generally has led to slightly less redundancy in amplifying the same

nucleotide sequence but also sampled slightly fewer nucleotide sequences in the group.

The invention provides a method of selecting a subset of primers sufficient to amplify all of a group of related nucleotide sequences. However, if desired, the subset of primers need not amplify all members of the group of related nucleotide sequences. Thus, the number of repetitions of the process of identifying any nucleotide sequences not amplified, or amplified only once by the set of primer pairs, can be limited. For example, the identified subset of primers can amplify a desired percentage of the group of related nucleotide sequences, which generally will be at least 80% of the group of nucleotide sequences, but can be 90%, 95% or 98% of the group, particularly 99% of the group.

The number of related nucleotide sequences in a group can vary. For example, a group can contain about 10 or more, about 20 or more, about 30 or more, about 40 or more, and generally about 50 or more, particularly about 75 or more or about 100 or more related nucleotide sequences. A large group of related nucleotide sequences can contain, for example, greater than about 100 sequences, generally greater than about 200 sequences, and particularly greater than about 400 sequences.

In the case where a large group of nucleotide sequences is of interest, the run time required for the computer process to identify a set of primers that amplify all of the nucleotide sequences can exceed a desirable length of time. Therefore, if desired, the computer process can be limited to a specified number of repeats of the process of identifying any nucleotide sequences not amplified, or amplified only once by the

set of primer pairs. However, even when sampling only a portion of the nucleotide sequences, the method of the invention provides an advantage over the previously used arbitrary primers, since a subset of such primers is obtained.

The subset of primers that amplify a group of related nucleotide sequences can be a minimal or near minimal set. Because the computer process has the ability to systematically identify all possible primers of a given size or range of sizes and match those primers with all of the group of related nucleotide sequences, a minimal set of primers containing the fewest number of primers sufficient to amplify all of the group of related nucleotide sequences can be selected as the subset of primers. However, if desired, the subset of primers need not be the minimal set. The subset of primers can be a near minimal set. For example, the identified subset of primers can be a near minimal set that is 20% more than the minimal number, generally 10% more than the minimal number, particularly 5% more than the minimal number. However, in some cases, it is not practical or desirable to determine the minimal set of primers sufficient to amplify all of a group of related nucleotide sequences. In this case, a desirable subset of primers is one in which a reduction in the maximal number of primers sufficient to amplify a group of nucleotide sequences is achieved using methods of the invention. Two primers, which can be different primers, are required to generate a PCR product. Therefore, the maximum number of primers sufficient to amplify a group of related nucleotide sequences is twice the number of related nucleotide sequences in the group. In some cases, a desirable subset of primers is one that is sufficient to amplify a group of related nucleotide sequences and contains about

50% of the maximum number of primers sufficient to amplify a group. In the case of a subset of primers containing 50% of the maximum number of primers sufficient to amplify a group, the number of primers in the subset is less than or equal to the number of related nucleotide sequences in the group. In addition, a desirable subset of primers can be one that is sufficient to amplify a group of related nucleotide sequences and that contains about 25% of the maximum number of primers, and particularly about 10% of the maximum number of primers.

To identify a subset of primers that amplify a group of related nucleotide sequences, the computer process selects a predetermined number of top ranked primers that match related nucleotide sequences in a group. The set of top ranked primers may not include one or more primers that, although not occurring in the largest number of nucleotide sequences, is required to select the minimal set of primers to amplify all of a group of related nucleotide sequences. Consideration of all primers that match a group of related nucleotide sequences rather than the top ranked primers is required to assure that the subset of primers selected is the minimal set. However, the run time required for the computer process to select the minimal set of primers sufficient to amplify all of the group of related nucleotide sequences can exceed a desirable length of time. Therefore, if desired, the computer process can be limited to a specified number of repeats of the process of identifying any nucleotide sequences not amplified, or amplified only once, by the set of primer pairs and generating a new set of top ranked primers. However, even when the subset of primers is a near minimal set, or even when the subset contains more than 50% of the

maximal number of primers sufficient to amplify a group of related nucleotide sequences, the method provides an advantage. The number of primers required to amplify a group of related nucleotide sequences is reduced significantly over the use of two primers for each nucleotide sequence, since the method identifies primer pairs that amplify multiple nucleotide sequences.

The invention also provides a method of using a set of primer pairs, which amplify an originally identified group of related nucleotide sequences, to amplify a population of nucleotide sequences related to the original group of nucleotide sequences. The primers in the selected primer subset, which are identified using a method of the invention, are synthesized by standard oligonucleotide synthesis techniques, which can be automated synthesis on a DNA synthesizer, such as an Applied Biosystems DNA synthesizer, or can be manual synthesis. The synthesized primers are useful reagents for amplifying a population of nucleotide sequences related to the original group of related nucleotide sequences. For example, the primers listed in Table III can be synthesized and used to amplify a population of nucleotide sequences related to the human nuclear receptor family. Alternatively, the primers listed in Table III can be synthesized with additional nucleotide bases added to one or both ends. The amplified nucleotide sequences are used to identify nucleotide sequences related to the original group of human nuclear receptors shown in Table II. Identification of the amplified nucleotide sequences is useful for characterizing a specimen with respect to a given set of conditions. As used herein, the term "specimen" refers to any biological material of interest, such as cultured cells, a cell lysate, or a whole organism, which can be

prokaryotic or eukaryotic, for example, cultured insect cells or mammalian cells or whole organisms such as mice or humans. For example, the expression of members of the human nuclear receptor family can be assessed using the nucleotide sequences amplified with the primers of Table III. Thus, the expression of human nuclear receptors, upon differentiation of P19 cells treated with retinoic acid, can be examined. The expression of human nuclear receptors also can be examined in tissue specimens taken at different stages of development, such as during development of a neonatal human to an adult. Accordingly, a subset of primers that amplify a group of related nucleotide sequences, as identified by a method of the invention, is useful as a reagent for examining expression of members of a group of related nucleotide sequences.

The disclosed method generates a subset of primer pairs that amplifies a group of related nucleotide sequences, for example, a set of primers that amplifies the human nuclear receptor gene family. Furthermore, the disclosed method provides a subset of primers that amplifies a population of nucleotide sequences and identifies amplified nucleotide sequences related to the original group of related nucleotide sequences. The amplified nucleotide sequence is a member of the group of related nucleotide sequences or is not a member of the original group of related nucleotide sequences. Thus, the invention provides a method to determine expression of a group of related nucleotide sequences, for example the expression of the related nucleotide sequences after exposure of cells to a drug or chemical agent. The invention also provides a method to systematically select a subset of primers to isolate and identify new nucleotide sequences, such as identifying nucleotide

sequences related to the human nuclear receptor gene family or the human G-protein coupled receptor gene family.

As exemplified herein, DNA sequence databases were used to generate a list of genes or partial gene sequences classified as members of the human nuclear receptor gene family (see Table II), the human G-protein coupled receptor gene family (see Table V), human apoptosis-associated genes (see Table VII) and human DNA repair and replication genes (see Table IX). Duplications in the list of genes or partial gene sequences can be removed. As used herein, the term "duplication" refers to a region of nucleotide sequence common to more than one nucleotide sequence with a predetermined level of identity. For example, the shorter of two sequences that had a predetermined identity of being at least 95% identical were removed from the list of genes or partial gene sequences of the human nuclear receptor family (see Table II), the human G-protein coupled receptor gene family (see Table V), human apoptosis-associated genes (see Table VII) and human DNA repair and replication genes (see Table IX). The shorter of any two sequences that overlap, where the overlap exceeds at least a predetermined identity, were also removed. As disclosed herein, a list of 44 human nuclear receptor genes or partial gene sequences was used as a group of related nucleotide sequences for identification of a set of primers that can amplify all members of the identified group of human nuclear receptor genes (see Example II). In addition, a list of 113 human G-protein coupled receptor genes or partial gene sequences was used as a group of related nucleotide sequences for identification of a set of primers that can

amplify all members of the identified group of human G-protein coupled receptor genes (see Example III).

A group of related nucleotide sequences that are structurally related, for example, the human nuclear receptor gene family and the human G-protein coupled receptor gene family, will have many candidate primer pairs overlap with conserved structural regions. Previously, degenerate primers derived from back translation of conserved amino acid motifs have been used to find new members of gene families (Carlberg et al., Mol. Endocrinol. 8:757 (1994)). The invention provides a method to identify related but unidentified members of a group of related nucleotide sequences, which also contain the conserved structural regions that define the family. Using sets of primer pairs that include these conserved regions allows the amplification of all, or nearly all, related nucleotide sequences. From this pool of amplified sequences, the sequences of unknown family members containing the conserved region can be identified. For example, RNA samples are isolated from different tissues of an organism, in this case a human, and used as specimens to amplify a population of nucleotide sequences related to the human nuclear receptor family or the human G-protein coupled receptor family. The amplified population of nucleotide sequences is cloned into a plasmid capable of replicating in bacteria, such as pBluescript (Stratagene; San Diego CA), and cloned nucleic acid molecules are sequenced using standard techniques to identify nucleotide sequences in the population that were not members of the original group of related nucleotide sequences.

The method can also be used to amplify sequences that are related by a common feature, such as

genes induced in response to a drug or chemical agent, genes involved in apoptosis, or genes involved in DNA repair and replication (see Examples IV and V). For example, a group of related nucleotide sequences can be identified that are genes induced upon stimulation of cells with retinoic acid. The group of nucleotide sequences related by a common feature can contain members that are not structurally related. For example, treating P19 cells with retinoic acid causes the cells to differentiate into neurons and glial cells and induces genes such as Mashl, a basic helix-loop-helix protein, EGF receptor, a cell surface hormone receptor tyrosine kinase, and choline acetyl transferase. While these genes share the common feature of being induced by retinoic acid, the proteins encoded by these genes carry out completely different functions in different compartments of a cell.

Functional protein domains are encoded by regions of nucleotide sequence that are conserved between structurally related proteins. As used herein, the term "functional domain of a protein" refers to a region or a portion of a protein that allows the protein to perform its biological purpose, such as an enzymatic domain, a structural domain, or any domain used by the protein to carry out its biological function. Primers that match with regions of a nucleotide sequence encoding such conserved domains also can amplify nucleotide sequences encoding structurally related family members. For example, primers that match with regions of the conserved kinase domain of the EGF receptor can amplify other tyrosine kinases. Therefore, to assure that the subset of primers selected exhibit specificity for the original group of related nucleotide sequences, regions of nucleotide sequence encoding conserved functional protein

domains can be excluded from the list of all primers. These regions of nucleotide sequence encoding conserved functional domains can be added, for example, to a list of abundant nucleotide sequences to be excluded from the list of primers. As such, a method of the invention can provide a subset of primers that amplifies a group of related nucleotide sequences in which all members are not structurally related.

Alternatively, an amplified population of related nucleotide sequences that share a common feature but are not all structurally related is a source from which structurally related sequences can be obtained. Proteins generally contain multiple functional domains, some of which are more highly conserved than others. For example, while the EGF receptor shares significant identity with the insulin receptor in its tyrosine kinase domain, which functions in intracellular signaling, the ligand binding domains do not share significant identity. Excluding primers that match with highly conserved functional domains, such as the kinase domain of the EGF receptor or the basic helix-loop-helix domain of Mash1, allows selection of a subset of primers that amplifies nucleotide sequences containing conserved structural domains. These nucleotide sequences containing conserved structural domains are related to an original group of related nucleotide sequences that are not all structurally related. For example, a subset of primers is selected that excludes the basic helix-loop-helix domain of Mash1 from the subset of primers that amplifies Mash1, EGF receptor and choline acetyl transferase. Such a subset of primers is used to amplify populations of nucleotide sequences in undifferentiated and differentiated P19 cells that contain members of the Mash1 family but not other nucleotide sequences that

share the basic helix-lop-helix domain. Thus, a subset of primers is selected that allows identification of nucleotide sequences that are structurally related to, but not members of, an original group of nucleotide sequence that are not all structurally related.

In addition, a group of related nucleotide sequences can be related by a common biological function such as being induced, for example, by apoptosis. In such a case, members of the group of related nucleotide sequences will include nucleotide sequences that are structurally related as well as nucleotide sequences that are not structurally related. For example, a model system such as HaCaT cells can be used to determine if a given set of genes of interest are expressed. Furthermore, apoptosis can be induced in HaCaT cells by the addition of sulindac (Hanif et al., Biochem. Pharmacol. 52:237-245 (1996); Lu et al., Proc. Natl. Acad. Sci. USA 92:7961-7965 (1995), each of which is incorporated herein by reference). Apoptotic cells such as sulindac treated HaCaT cells can be used to identify genes induced by apoptosis or as a model system to confirm that a given subset of primers can amplify genes of interest induced by apoptosis.

In another embodiment, primer pairs can be selected to perform PCR fingerprinting of a specimen of interest. The computer process can require that primer pairs generate PCR products that differ by at least a predetermined size range such as ± 3 base pairs. Thus, PCR products of sizes that are impractical to separate using predetermined analytical techniques are not generated. For example, in the case of related gene families such as the human nuclear receptor family, the distance between conserved structural regions is often

conserved. Primers from these conserved regions, therefore, are predicted to generate similar sized PCR products. Eliminating these primers from the matrix maximizes the generation of dissimilar sized PCR products, which can be used to fingerprint a sample with respect to a given set of genes. Thus, the invention provides a means to perform PCR fingerprinting.

A particularly valuable use of PCR technology is PCR fingerprinting. As used herein, "PCR fingerprinting" refers to using PCR to generate a set of nucleotide products of differing sizes that can be used to discriminate between a predetermined set of specimens. PCR fingerprinting has been used to detect polymorphisms in related genomes (Welsh et al., Nucleic Acids Res. 19:303 (1991); Welsh and McClelland, Nucleic Acids Res. 19:861 (1991); Woods et al., J. Clin. Microbiol. 31:1927 (1993)). PCR fingerprinting also has been used to study differential expression of arbitrarily sampled RNAs (Welsh et al., Nucleic Acids Res. 20:4965 (1992); Liang and Pardee, Science 257:967 (1992)). In addition, PCR fingerprinting has been used to enrich for genes of interest (Stone and Wharton, Nucleic Acids Res. 22:2612 (1994); Yoshikawa et al., Biochim. Biophys. Acta 1264:63 (1995)).

The products of a PCR fingerprinting reaction also can be used as probes for differential hybridization. The products of a PCR fingerprinting reaction are enriched for the related nucleotide sequences of interest. These PCR fingerprinting products will be effective as probes in differential hybridization experiments examining the related nucleotide sequences of interest because the complexity of the probe is much lower than for total cDNA. In another example of PCR

fingerprinting, primers can be designed to generate PCR fingerprints that can be used to quantitate the level of a particular nucleotide sequence relative to other nucleotide sequences within the group of related nucleotide sequences.

In another embodiment, a method of the invention is used to isolate 3′ ends of mRNA sequences, if the 3′ ends are known. Anchored oligo-(dT) primers have been used with arbitrarily selected primers or with a primer selected to match perfectly in one known gene to amplify the 3′ end of that gene (Liang and Pardee, *supra*, 1992). As used herein, the term "anchor primer" means a primer that can amplify the 3' end of an mRNA, generally by hybridizing to the poly(A) tail of an mRNA. As such, an anchor primer will generally be an oligo(dT) primer comprising a nucleotide sequence of the formula T_n, where n designates the number of Ts in the primer. A method of the invention, which can be a computer process, is used to identify a smaller number of primers than the number of related nucleotide sequences in a group or even a single primer that hybridizes to the opposite strand relative to the anchored oligo-(dT) primer. This identified primer with the anchored oligo-(dT) primer can generate PCR products from a known set of mRNA sequences, allowing one primer pair to amplify the 3′ ends of multiple mRNA sequences.

The use of anchored primers to sample the 3' end of mRNAs has several advantages. For example, where EST libraries are being probed, many EST libraries are naturally 3' biased so a probe of a given complexity that is also 3' biased should hybridize to more clones than a probe derived from internal sampling of mRNAs. In contrast, when using two arbitrary primers, many of the

resulting PCR products will be derived from the middle or 5' end of mRNAs. These internal and 5' products are less likely to occur in a sample that is 3' biased, for example, a 3' biased EST library and such internal and 5' products might contribute to the background. Thus, with a 3' biased probe, the overall probe complexity can be lower for any given throughput of positive signals. However, when screening randomly primed EST libraries, internal priming is more desirable.

Another potential advantage of 3' sampling is that the 3' ends are generally the most divergent part of a mRNA. Thus, if there are closely related family members in a group of related nucleotide sequence that need to be distinguished, this distinction of closely related members is most easily achieved in the 3' non-coding region of the mRNA.

Sampling of 3' ends of mRNAs is useful depending on the intended use of the PCR products. For example, because an oligo(dT) primer can prime in multiple registers on the poly(A) tail, the PCR products can vary in size and not generate a discrete fingerprint. However, when the PCR products are intended to be used as probes, variable sizes in PCR products is not problematic. Therefore, an oligo(dT) primer combined with one or more specific primers, which can amplify more than one member of a group of related nucleotide sequences, can be used to generate a probe.

When it is desirable to use the PCR products to obtain a fingerprint of a group of related nucleotide sequences, anchor primers that amplify the 3' end of a nucleotide sequence can be still be used. For example, it is convenient to determine a fingerprint of a sample

by analyzing the PCR products on a gel. Such a gel analysis, which is convenient, fast, and economical, can be used to determine the quality and reproducibility of the probe at multiple RNA concentrations before using it as a probe.

Amplifying the 3' ends of nucleotide sequences using oligo(dT) primers has generated useful fingerprints. For example, oligo(dT) probes such as T_nV, where V is G, C or A, have been used successfully to generate stable fingerprints. Additional oligo(dT) probes that have been used to successfully generate fingerprints include T_nG, T_nA, and T_nAC.

Although sampling the 3' end of nucleotide sequences can be advantageous as described above, potential disadvantages of 3' anchoring can occur if 3' selectivity at the end of the oligo(dT) is not good, which can result, for example, in overlap in genes hybridized by fingerprints primed with different anchored oligo(dT) primers. This disadvantage has not been observed for the prominent products seen on a gel, which correspond to more abundant sequences in the sample, but it can be a problem for less abundant nucleotide sequences.

An additional disadvantage of sampling the 3' end of a nucleotide sequence can occur if a nucleotide sequence in the sample has a region that is difficult to amplify close to the poly(A) tail. However, it is likely that only a very few mRNAs will be inefficiently amplified due to structures at the 3' end. Alternatively, these disadvantages can be overcome, if necessary, by using internal primers, or primers that

amplify the 5' end of a nucleotide sequence, as described herein.

The invention also provides a subset of primers sufficient to amplify a group of related nucleotide sequences, wherein the subset comprises at least one anchor primer of the formula T_nX_m, wherein X is selected from the group consisting of G, A, C and T, n is a number between 10 and 20 and m is a number between 0 and 3; and wherein the subset comprises one or more second primers, wherein the second primer combined with the anchor primer amplifies two or more related nucleotide sequences in the group. In such a case, the second primer is not an anchor primer. The anchor primer can also have the formula T_nX_m, where n is a number between 5 and 50 and m is a number between 0 and 10.

In another embodiment, a method of the invention is used to identify degenerate primers. As used herein, the term "degenerate primer" refers to a primer sequence that contains at least one position, X, that has more than one nucleotide, where X is A, G, T, C or a modified base such as inosine (I). For example, in a set of primers containing the degenerate 8-mer primer GATXCCGT, a set of primers contains GATACCGT and GATTCCGT. In another example, in a set of primers containing the degenerate 8-mer primer CATCXAGG, the set of primers contains CATCGAGG, CATCTAGG, CATCCAGG, CATCIAGG. The statistical frequency of matching a given nucleotide sequence using degenerate primers increases with the number of primers in the set representing that degenerate primer, in these examples 2 and 4 primers, respectively, representing the degenerate primer. The method provides an advantage in that longer primers,

which would normally occur at low frequency in a group of related nucleotide sequences, can be identified.

The use of degenerate primers allows, for example, some primers 10 or more bases long to occur many times in small groups of related nucleotide sequences. An example would be primers of the form XXZXXZXXZXX, where each X is a different specified base and Z is either R (purine) or Y (pyrimidine). An example of such a primer is GAYTCRTCYCC. The potential advantage of using such a degenerate primer is that the primer can hybridize to nucleotides encoding clusters of amino acid that are common among members of a group of related nucleotide sequences, even if the group is not from phylogenetically related genes. For example, it is known that the motif RNYRNYRNY is common in all open reading frames, perhaps as a vestige of the original genetic code in primordial organisms. Degenerate primers can therefore be advantageously used to obtain longer primers that will match to a group containing a smaller number of related nucleotide sequences.

In another embodiment, a method of the invention is used to sample a large number of nucleotide sequences. Public databases, such as the THC database of the Institute for Genome Research, contain in excess of 50,000 nucleotide sequences from the 3' ends of human mRNA sequences. Practical limitations to a predetermined analytical technique can require that limits be placed on the number of PCR products generated when large numbers of nucleotide sequences are to be analyzed. Therefore, minimizing the number of PCR products generated from an individual nucleotide sequence can be desirable.

A subset of primer pairs, where each primer pair generates 40 to 60 PCR products from a large group of related nucleotide sequences such as human mRNA sequences, is identified using the computer process in order to minimize the number of PCR products generated from an individual nucleotide sequence. Using this subset of primer pairs, a matrix is generated that ranks the nucleotide sequences with respect to the number of primer pairs that can amplify a nucleotide sequence. The nucleotide sequence amplified by the fewest number of primer pairs is selected. Among the primer pairs that amplifies the first selected nucleotide sequence, the primer pair that amplifies the most other nucleotide sequences (for example, up to 60 nucleotide sequences) is selected. All nucleotide sequences amplified by this selected primer pair are removed from the matrix, for example, up to 60 nucleotide sequences are removed. A new matrix is generated that ranks the remaining nucleotide sequences with respect to the number of primer pairs that can amplify a nucleotide sequence. The new matrix will contain, for example, up to 60 fewer nucleotide sequences than in the first matrix. The nucleotide sequence remaining in the matrix that is amplified by the fewest number of primer pairs is selected. A primer pair that amplifies the second selected nucleotide sequence and that amplifies the most other nucleotide sequences (for example, up to 60 nucleotide sequences) is selected. All nucleotide sequences amplified by this selected primer pair are removed from the matrix. This process is repeated until a subset of primer pairs is identified that amplifies all of the related nucleotide sequences. This subset of primer pairs minimizes redundant sampling of the nucleotide sequences in the original list due to removal of large numbers of nucleotide sequences, for example, up

to 60 nucleotide sequences, at each iteration of the process. This subset of primer pairs produces the fewest number of PCR products from any single nucleotide sequence. An identical procedure, selecting a different starting primer pair that amplifies the first selected nucleotide sequence, can be used to generate a different subset of primer pairs that will also minimize redundant sampling of the original list of nucleotide sequences.

In another embodiment, the 3′ ends of nucleotide sequences derived from mRNA sequences can be isolated using anchored primers such as oligo (dT)C, oligo (dT)G, or oligo (dT)A (Liang and Pardee, *supra*, 1992). The computer process is used to identify primers that would amplify, with one of the anchor primers, a predetermined range of PCR products, for example, 40 to 60 PCR products, from a data base such as a database containing 3′ expressed sequence tags (ESTs). When all three anchor primers are used with the identified primer, 120 to 180 PCR products are generated. The identified primer, with the anchor primers, amplifies the 3′ ends of mRNA sequences of the group of related nucleotide sequences as well as previously unknown related nucleotide sequences not present in the currently known database of ESTs.

Identifying primer pairs capable of amplifying all nucleotide sequences from a large group of nucleotide sequences requires generation of a very large number of PCR primers. The invention provides a method to identify a subset of primer pairs that amplifies all of the sequences in a large group of related nucleotide sequences, while minimizing the number of primer pairs sufficient to amplify all of the nucleotide sequences. For example, using a database containing a large group of

nucleotide sequences, such as 50,000 mRNA sequences, primers are identified that can generate a predetermined number of PCR products, such as 40 to 60 PCR products, when paired with one of the anchor primers. A matrix is generated that ranks the large group of nucleotide sequences, 50,000 in this example, with respect to the number of primer pairs that can amplify a nucleotide sequence, where the primer pairs are derived from the identified primers and the anchor primers that generate 40 to 60 PCR products.

To minimize redundant sampling of the nucleotide sequences, the nucleotide sequence amplified by the fewest number of primer pairs is selected. Among the primer pairs that amplifies the first selected nucleotide sequence, the primer pair that amplifies the most other nucleotide sequences (up to 60 nucleotide sequences) is selected. All nucleotide sequences amplified by this selected primer pair are removed from the matrix, in this example up to 60 nucleotide sequences are removed from the matrix. A new matrix is generated that ranks the remaining nucleotide sequences with respect to the number of primer pairs that can amplify a nucleotide sequence. The process of identifying nucleotide sequences amplified by a primer pair and removing those identified nucleotide sequences from the list is repeated until all nucleotide sequences have been removed from the list and a set of primer pairs has been generated that can amplify all nucleotide sequences in the list. Because this set of primer pairs minimizes redundant sampling of the nucleotide sequences in the original list, the number of primers sufficient to amplify all of the nucleotide sequences has been minimized. In this example using 50,000 nucleotide sequences, a set of about 1500 primers can be identified

that amplifies the entire list of 50,000 nucleotide sequences.

Information is rapidly accumulating on ESTs (expressed sequence tags), which can also be used to identify abundant mRNAs. For example, a rapidly increasing number of ESTs are available in databases, including public databases such as GenBank (http://www.ncbi.nlm.nih.gov/ dbEST/ index.html). A list of ESTs compiled from available databases can be used to exclude primers that match those ESTs that are highly expressed in many cell types. Furthermore, such a list of ESTs can be used to develop primer sets that will allow the sampling of virtually all known ESTs or a user specified subset of ESTs.

Previously, primer sets that selectively amplified large sets of genes were determined by randomly pairing a set of primers using a Monte Carlo method (Lopez-Nieto and Nigam, Nature BioTechnology 14:857 (1996)). The primer sets were selective for protein-coding regions of structurally related genes.

A method of the invention provides a systematic approach to identifying primer pairs that can sample a group of related nucleotide sequences, using a variety of predetermined criteria. The group of related nucleotide sequences need not be structurally related. The method disclosed herein systematically maximizes the number of nucleotide sequences that can be amplified with a set of primers by systematically pairing primers using a matrix. The invention disclosed herein provides a method to impose specified criteria on primers and on the PCR products generated by those primers. For example, the method allows imposing exclusion of primers that match

undesirable nucleotide sequences, requiring a minimum number of PCR products be produced by a primer pair, exclusion of primers that match conserved structural regions that encode functional protein domains, and requiring that primer pairs generate different sized PCR products. In addition, a method of the invention disclosed herein allows generating primers that match to any region of an mRNA sequence, not just the coding region.

Primer sets that sample open reading frames (ORF) in mammalian mRNA sequences have been described (Lopez-Nieto and Nigam, *supra*, 1996). The ORF-specific 8-mer primers sampled 14% of the human nuclear receptor genes, 28% of the human G-protein coupled receptor genes, 40% of the human apoptosis-associated genes and 31% of the human DNA repair and replication genes identified as groups of related nucleotide sequences in the Examples herein. In contrast, as disclosed herein, the methods of the invention provided identification of sets of primer pairs that sampled 100% of the human nuclear receptor genes and human G-protein coupled receptor genes as well as 93% of the human apoptosis-associated genes and 98% of the human DNA repair and replication genes. If abundant nucleotide sequences are removed from these previously identified ORF-specific 8-mer primers, only 3 of the 113 human G-protein coupled receptor genes are sampled (see Table V). Similarly, a set of 10-mer primers that sample mammalian G-protein coupled receptor genes was also identified (Lopez-Nieto and Nigam, *supra*, 1996). When the previously identified 10-mer primers matching abundant nucleotide sequences were removed, none of the 113 human G-protein coupled receptor genes were sampled. Therefore, the invention provides a substantial advantage

over previous methods, such as that reported by Lopez-Nieto and Nigam (Lopez-Nieto and Nigam, *supra*, 1996).

The invention also provides a subset of primers of sufficient to amplify a group of related nucleotide sequences comprising nuclear hormone receptor genes. A subset of such primers can be selected from the group consisting of TGCAAGGG; TGCAGGAG; CAGCAGCG; GGCTGCAA; GCCTCCAG; TCCTGGAG; CTGCCTGG; CCTTCCTC; CTCCCTGG; CTGCCCTG; AGGGCTGC; CTGCTGGA; CCGCTGCC; GGAGGCAG; AGCCTGGA; GGGCAGAG; GGCAGCTG; GAGGAAGG; CAGCTGCC; GATTCCAC; GATGAGCT; CTTCTGGA; and CTGGAGCT. The invention also provides the primers shown in Table III.

The invention additionally provides a subset of primers sufficient to amplify a group of related nucleotide sequences comprising G-protein coupled receptors. A subset of such primers can be selected from the group consisting of GCTGGCCA; TCTGCTGG; CTGTGCTG; TGCTGGCT; ACCTGCTG; CTGCTGGT; CCTGGCCT; TCATCTGC; CTTCCAGG; TCCTGGCC; GCTGGCTG; GGCCCTGG; TGCCCTGG; GCTGGTGG; CTCTGCTG; ACAGCAGC; GCTGCCTC; CCAGGGCT; GCTGCCCC; TGGCCATT; GCCCTGGG; CCACCAGC; TCCTGCTG; GGCCACCA; TGGTGGCC; CTGCTGCT; CTCCTTCT; TCCTGGTG; CTGGGCCA; GGCAGCTG; TGCTGGGC; CTGCTGTG; GCCTCTGC; CTGGCCAG; CTGGCCAC; CTGCCTCC; TGTGGCCC; GGCTATGT; TCCAGTCC; TGGCCAGC; CAGCACAG; CAGCAGCG; and CAGCCAGC. The invention also provides the primers shown in Table VI.

The invention further provides a subset of primers sufficient to amplify a group of related nucleotide sequences comprising apoptosis-associated genes. A subset of such primers can be selected from the group consisting of CTGGAGGA; TCATCCAG; CTGGAGAA;

GCTGCAGC; CTGCTGGA; GAACAGGA; GCTCCTGG; GCCCCTGG; CCAGAGCA; CAGCCCAG; CTTCCTGA; TGGATGCA; TCCAGTTC; TGAAGAGC; CCTGGGAG; CTCCCAGG; CCAGGCAG; CAGAGGAG; CCTCCAGG; GGAGGCAG; GGCGGCTG; TCTGCTGG; GCTCGGCC; GCTGGTGG; GGCAGCTG; TCCTGGGT; CTGCCTCC; CTCAGCAG; GAACTGGA; CAGCTGGA; CAGCCGCC; CTGCATCC; TGCAACAG; GGCTGCGC; CTGGCCAG; GTGGTGGA; CCTGCAGG; and GCCTCCAG. The invention also provides the primers shown in Table VIII.

The invention additionally provides a subset of primers sufficient to amplify a group of related nucleotide sequences comprising DNA repair and replication genes. A subset of such primers can be selected from the group consisting of GGAAGGAG; TGCAGGAG; CTGGCTGA; CTTCCTCA; TCATCCAG; AGCAGCAA; AGGCTGGG; AGAAGGAG; CTGGAGGA; AGCTGAAG; CATCCAGA; GCTGCAGC; CTTCCTGA; CCTCCTGG; TGCTCTGG; CTGCTGAA; GCTGCTGA; TCCAGAAG; GCAGCTGC; CTGCTGGA; GGCAGAGG; AGTCTGGA; AGTTTGCC; GTCCAGGT; CACAGCTG; TTGCTGAC; CTCAGTGC; TGCCACCA; AGAACCTG; ACCATTGC; GGAGGCAG; CCAGGAGG; TGGAGAGA; CTGATGAC; GAGATGGA; AGATGCTG; GCTGGAAG; GCCAGGAG; CAGGAAGC; TCTGGAAG; CATCTGGA; and TGAGGAAG. The invention also provides the primers shown in Table X.

The invention also provides a computer apparatus comprising a processor, main memory in communication with the processor, and a primer pair selector in communication with the main memory for carrying out the computer-executed steps of identifying a group of related nucleotide sequences; generating a set of primers that matches each of the related nucleotide sequences; determining for each systematic pairing of each primer which of the related nucleotide sequences are amplified; and selecting from the systematic pairings a

subset of primers sufficient to amplify the group of related nucleotide sequences.

The invention also provides a computer program product for determining a set of primer pairs sufficient to amplify a group of related nucleotide sequences comprising means for identifying a group of related nucleotide sequences; means for generating a set of primers that match each of the related nucleotide sequences; means for determining for each systematic pairing of each primer which of the related nucleotide sequences are amplified; means for selecting from the systematic pairings a subset sufficient to amplify the group of the related nucleotide sequences; and signal-bearing media containing the means for the identifying, generating, determining and selecting.

Referring to Figure 1, a block diagram of computer system 10, which is preferably employed to implement the present invention, is shown. Computer system 10 has operating system 11, processor 12, main memory 14, primer pair selector 16, display screen 20, input device 22, media drive 24, and disk storage 26, each of which is connected to system unit 10. Operating system 11 is an operating system such as UNIX, MS-DOS, Windows, or OS/2. In the preferred embodiment, the processor 12 is a general purpose programmable processor such as an Intel PENTIUM processor or a Motorola 68,000 processor, suitable for a mid-size computer such as DEC or IBM. The main memory 14 can be well known random access memory (RAM) that is sufficiently large to hold the necessary programming and data structures. The primer pair selector 16 in communication with main memory carries out the computer-executable steps of identifying a group of related nucleotide sequences; generating a set

of primers that matches each of the related nucleotide sequences; determining for each systematic pairing of each primer, which of the related nucleotide sequences is amplified; and selecting from the systematic pairings a subset sufficient to amplify all of the related nucleotide sequences. The display screen 20 is a screen for visualizing, for example, input data. The input device 22 is a mouse or a keyboard, or a combination thereof, or any other device to input information. The media drive 24 is a drive, such as a tape drive, a disk drive or a CD drive, that provides the computer system 10 access to the primer pair selector 16. The disk storage 26 is a device, such as magnetic tape or Zip disk, that provides storage capacity for data.

Referring to Figure 2, the computer-executed steps for carrying out the method of the invention by the primer pair selector 16 is shown. Step 100 starts the implementation of the present invention. In step 102, a group of related nucleotide sequences is identified. In step 104, a set of primers that matches each of the group of related nucleotide sequences is generated. In step 106, the group of related nucleotide sequences amplified by a primer pair is determined for each systematic pairing of each primer. In step 108, from said systematic pairings, a subset of primers, sufficient to amplify all of the related nucleotide sequences, is selected. The method steps related to selecting a subset of primers sufficient to amplify a group of related nucleotide sequences end in step 110.

Referring to Figures 3 and 4, the flowcharts that depict the computer-executed steps for carrying out the method of the invention are shown. Referring to Figure 3, step 200 starts the implementation of the

present invention. In step 202, a group of related nucleotide sequences, referred to as sequences in Figures 3 and 4, is identified. If desired, duplications can be removed from a list containing the group of related nucleotide sequences. For example, the shorter of two sequences that have greater than 95% identity can be removed from the list. In step 204, a list of all possible primers is compiled. The list of primers can be constrained to predetermined criteria, for example, a window of G+C content can be imposed, primers can be limited to a specified length or range of lengths, or any combination of these and other criteria can be imposed. In step 206, the primers are ranked by the number of nucleotide sequences that the primer is capable of sampling. In step 208, a set of top ranked primers is generated by selecting the top ranked primers and adding the top ranked primers and their complements to the set. The top ranked primers are those that sample the largest number of nucleotide sequences. For example, the top 30 primers can be selected, generating a list of 60 primers, including the selected primers and their complements. The list of primers can be constrained to predetermined criteria, for example, primers hybridizing to undesirable nucleotide sequences can be excluded.

In step 210, each primer in the top ranked primer set is paired to all primers in the set. In step 212, the related nucleotide sequences which are sampled by primer pairs are determined for each primer pair. If desired, additional constraints can be imposed on the primer pairs. The computer process can require that the PCR products be limited to a predetermined size, for example greater than 100 base pairs or less than 1000 base pairs. Furthermore, the computer process can require that primer pairs generate PCR products that

differ by a predetermined size range, such as ± 3 base pairs. Thus, PCR products of sizes that are impractical to separate using predetermined analytical techniques will not be generated. Primer pairs also can be limited to those primer pairs that generate a minimum number of different sized PCR products such as three or more different sized PCR products. In addition, the different PCR products can be limited to PCR products derived from different nucleotide sequences. In step 214, a matrix is generated that ranks the nucleotide sequences by the number of primer pairs that generate simulated PCR products and the primer pairs by the number of sequences sampled. In step 216, a nucleotide sequence is selected by a predetermined criterion. For example, the nucleotide sequence sampled by the largest number of primers can be selected.

Referring to Figure 4, in step 218, a primer pair that amplifies the selected sequence is identified. If desired, predetermined criteria can be imposed on the primer pairs. For example, the primer pairs can be required to generate PCR products of a specified range, such as greater than 100 base pairs or less than 1000 base pairs. In another example, the primer pairs can be required to sample a minimum number of nucleotide sequences, such as more than two different nucleotide sequences. In another example, the primer pairs can be required to generate PCR products that differ by a predetermined value, such as the PCR products generated must differ by at least ± 3 base pairs. In another example, if more than one primer pair amplifies the selected sequence, the primer pair that samples the selected nucleotide sequence and that samples the largest number of other nucleotide sequences in the group of related nucleotide sequences is identified. If desired,

any one of the above mentioned criteria, or any other criteria, alone or in any combination, can be imposed on the primer pairs.

In step 220, a subgroup is created by identifying all nucleotide sequences in the group of related nucleotide sequences that are amplified by the identified primer pair. In step 222, the identified primer pair and those nucleotide sequences in the subgroup that are sampled by the primer pair are removed from the matrix. In step 224, a new matrix is created that contains the remaining primer pairs and remaining nucleotide sequences.

In step 226, an inquiry is performed to determine if there are any nucleotide sequences remaining in the new matrix that are sampled by any primer pairs remaining in the new matrix. In step 228, if the answer is "yes", then a nucleotide sequence remaining in the new matrix is selected by the predetermined criterion (go to step 216). In step 230, if the answer is "no", a subset of primers from all primer pairs identified in step 218 that sample all of the group of related nucleotide sequences is selected. If desired, any nucleotide sequences that are not sampled by primer pairs derived from the set of primers selected in step 208 can be identified. These identified nucleotide sequences which are not sampled by the original set of top ranked primers can be used to generate a new set of primers that match the related nucleotide sequences that are not sampled. The new set of top ranked primers is generated by compiling a new list of all primers that match the nucleotide sequences not sampled. The top ranked primers in the new list and their complements are combined with the top ranked primers identified in the original list to

generate a new set of top ranked primers as in step 208. Again, any nucleotide sequences that are not sampled by the new set of top ranked primers can be identified and a new list of primers can be compiled. The process can be repeated until all of the nucleotide sequences in the original group of related nucleotide sequences are amplified. If desired, a limit to the number of repeats of identifying nucleotide sequences not sampled by the primers in the top ranked primer list can be imposed, such as limiting the number of repeats to three. The subset of primers selected by the method can amplify all of the group of related nucleotide sequences. However, if desired, the subset of primers need not amplify all members of the group of related nucleotide sequences. The number of repeats of the process of identifying any nucleotide sequences not sampled, or for example, sampled only once, can be limited such that the identified subset of primers amplifies a desired percentage of the group of related nucleotide sequences. The method steps related to selecting a subset of primers sufficient to amplify a group of related nucleotide sequences end in step 232.

Referring to Figure 5, a program product 700 for executing, in combination with the computer system 10, and for performing the method described above is shown. The program product 700 includes a computer readable medium 702, such as a floppy disk, readable by media drive 24 containing signals 704 to 710 recorded thereon. Identifier signals 704 are means for identifying a group of related nucleotide sequences. Identifier signals 706 are means for generating a set of primers that matches each of a group of related nucleotide sequences. Identifier 708 are means for determining, for each systematic pairing of each primer, which of the related nucleotide sequences are amplified.

Identifier 710 are means for selecting from the systematic pairings a subset of primers which can amplify all of the related nucleotide sequences. Thus, the medium 702 is signal bearing media containing said means for identifying, generating, determining, and selecting as described above.

It is understood that modifications which do not substantially affect the activity of the various embodiments of this invention are also included within the definition of the invention provided herein. Accordingly, the following examples are intended to illustrate but not limit the present invention.

EXAMPLE I

Selection of PCR Primer Pairs to Amplify a Group of Related Nucleotide Sequences

This example demonstrates the selection of a set of PCR primers that amplify a group of related nucleotide sequences.

A. Selecting an Upper Limit for Length of Primers that Match Members of a Group of Related Nucleotide Sequences

The program WORDUP was adapted to generate a list of the most common 9-mer, 8-mer and 7-mer primers from an identified group of related nucleotide sequences (Pesole et al., *supra* (1992)). The program was written in C language on a UNIX operating system. The appropriate upper limit of primer length was determined. A primer 10 nucleotides in length is expected to occur about once every 1,000,000 (4^{10}) base pairs. The most common 10-mers in a group of related nucleotide sequences were identified. If the 10-mers occurred only a few

times and were confined to regions encoding conserved protein domains, the primers were limited to 9-mers or shorter. For the groups of related nucleotide sequences used in these Examples, the primers were limited to 9-mers or shorter.

B. Removal of Primers that Match Abundant Nucleotide Sequences

A list of accession numbers of abundant nucleotide sequences was compiled. The list contained human ribosomal RNA, human mitochondrial DNA, Alu elements and mRNA sequences carrying fragments of LINE elements (Table I).

Every primer in the list of most common 9-mer, 8-mer and 7-mer primers was compared to the human ribosomal RNA, human mitochondrial DNA and human Alu element sequences and those primers that matched these abundant nucleotide sequences were removed from the list of primers. The remaining primers in the list were compared to the list of 99 mRNA sequences carrying fragments of LINE elements and those primers that occurred three or more times in the list of LINE elements were removed from the list of primers. The vast majority of 16,384 possible 7-mers occurred in the human ribosomal

Table I. GenBank Accession Numbers of Abundant Nucleotide Sequences

Ribosomal RNA and Mitochondrial DNA

K03432|HUMRGEA Human 18S rRNA gene
M11167|HUMRGM Human 28S ribosomal RNA gene
J01866|HUMRRB Human 5.8S ribosomal RNA
D38112|HUMMTA Human mitochondrial DNA, complete sequence

Alu elements:

U14567	U14568	U14569	U14570	U14571
U14572	U14573	U14574		

LINE Elements and 3' mRNAs fragments carrying parts of LINE elements:

AA017128	AA018943	AA022026	AA055654	AA057222
AA074788	AA076364	AA081957	AA081993	AA082639
AA084139	AA084303	AA085706	AA085707	AA088273
AA088381	AA095194	AA100276	AA102000	AA112088
AA112316	AA112323	AA113978	AA121767	AA121839
AA121840	AA121875	AA121876	AA121916	AA126794
AA126847	AA128621	AA128858	AA129985	AA130476
AA130536	AA131481	AA136629	AA136637	AA136721
AA136934	AA136977	AA148366	C17235	D58460
H03599	H13052	H20876	H82488	H85238
H92806	M78222	M85371	N20521	N22643
N23244	N23646	N23655	N23657	N23864
N24958	N25053	N29555	N33076	N41014
N70045	N76123	N76330	N79938	R14820
T02866	T02882	T03057	T03259	T06602
T06958	T07197	T16214	T48647	T56669
T57073	T57474	T57745	T59577	T60735
T63720	T90251	W03161	W03511	W19702
W26931	W26997	W27003	W37681	W58442
W85828	W90097	W90195	W93703	

RNA, mitochondrial DNA, Alu elements or LINE elements shown in Table I. Thus, primers were limited to those primers greater than seven nucleotides in length.

Primers eight or nine nucleotides in length were considered for further analysis. The typical 8-mer

has a frequency of about one in 65,000 (4^8) base pairs. The typical 9-mer has a frequency of about one in 262,000 (4^9).

C. Selecting a G+C Content of Primers

Consideration was given to the stability of primer and template interactions. For successful PCR, both primers in a primer pair need to be matched with regard to their melting temperature on the template. Primers that are A+T rich interact less strongly than G+C rich primers. Therefore, further analysis was confined to those primers that were 50% to 90% G+C.

D. Criteria for Selection of Related Nucleotide Sequences

A group of related nucleotide sequences was identified. Duplicate sequences with greater than 95% identity were identified and the longest of the nucleotide sequences in the identified group was retained in the group using the computer process CLEANUP (Grillo et al., Comput. Applic. Biosci. 12:1 (1996)), which is incorporated herein by reference. The shorter of the duplicate nucleotide sequences was removed from the list. Nucleotide sequences less than 800 bases in length were also removed from the list.

E. Systematic Pairing of Primers to Determine Related Nucleotide Sequences That Are Amplified

The remaining 8-mer and 9-mer primers were ranked in the order of the number of related nucleotide sequences to which the primers matched. The top 50 primers matching the largest number of related nucleotide sequences were selected and these 50 primers and their

complements were compiled into a list of primers. Simulated PCR products from 50 to 1000 base pairs were determined for each primer pair combination. A matrix was generated that ranked the related nucleotide sequences by the number of primer pairs that can generate a simulated PCR product and the primer pairs by the number of related nucleotide sequences that can be sampled.

The most frequently sampled related nucleotide sequence was selected as the first sequence to consider. The primer pair that sampled the first selected nucleotide sequence was used to identify the remaining related nucleotide sequences which the selected primer pair sampled. The primer pairs were required to sample at least two different nucleotide sequences. The primer pair and the first selected nucleotide sequence sampled by the primer pair were removed from the matrix. The most frequently sampled related nucleotide sequence remaining in the list was selected, and the primer pair that sampled this second selected related nucleotide sequence was selected and used to identify the remaining related sequences which the selected primer pair sampled. The primer pair and the identified nucleotide sequences sampled by the primer pair were removed from the matrix. The process of selecting a related nucleotide sequence and the primer pairs that generated simulated PCR products was repeated until a set of primer pairs was identified that sampled the entire list of related nucleotide sequences. If a first iteration of the above procedure did not yield a set of primer pairs that sampled all of the related nucleotide sequence, those related nucleotide sequences not sampled in the first matrix were compiled into a new list to repeat the

procedure. In these examples, the number of iterations was limited to two.

Criteria for the primer pairs and PCR products were applied. A group of related nucleotide sequences that are structurally related will have many of the most frequent candidate primer pairs from conserved structural regions. The spacing between nucleotide sequences that encode conserved protein motifs are often conserved in some structurally related nucleotide sequences. In this case, primers that originate in two conserved domains will yield similar sized products. Therefore, the maximum number of products of identical size for any primer pair was set to three or less to minimize the number of identically sized PCR products.

Additional criteria for the primer pairs and PCR products were applied. If more than one PCR product was found for a given nucleotide sequence, the shorter of the PCR products was chosen. Also, the minimum number of PCR products for each nucleotide sequence was set at three to maximize sampling of each nucleotide sequence. PCR products also were limited to those that have different PCR primers at each end.

These results demonstrate that a subset of PCR primers that amplifies a group of related nucleotide sequences can be selected using the disclosed method.

EXAMPLE II

Selection of PCR Primer Pairs to Amplify the Human Nuclear Receptor Family

This example demonstrates that a subset of PCR primers that amplifies the human nuclear receptor gene

family can be selected using the disclosed method. A computer process was used to generate a set of primer pairs that amplifies the human nuclear receptor gene family (see Example I and Figure 2).

Members of the human nuclear receptor gene family were identified (Mangelsdorf et al., Cell 83:835 (1995)), which is incorporated herein by reference. These genes encode proteins that interact with small signaling molecules such as retinoids and steroids and bind to transcriptional regulatory elements in the cell nucleus. A list of 62 mRNA and mRNA fragments from the human nuclear receptor gene family was compiled. Following removal of duplicates and nucleotide sequences less than 800 bases, 44 human nuclear receptor mRNA sequences remained in the list (see Table II).

Table II. Human Nuclear Receptor mRNA Sequences and mRNA Fragments[1]

1 D49728	2 J03258	3 L02932	4 L07592	5 L31785
6 L40904	7 L76571	8 X03635	9 M15716	10 M16801
11 M24748	12 M24857	13 M24899	14 M26747	15 M29960
16 M34233	17 M62760	18 U04897	19 U04898	20 U04899
21 U07132	22 U10990	23 U12767	24 U22662	25 U28386
26 U31929	27 U41065	28 U64876	29 U79012	30 U80802
31 X03225	32 X06538	33 X07282	34 X12794	35 X12795
36 X52773	37 X63522	38 X66424	39 X69068	40 X75918
41 X89894	42 Z30425	43 Z35491	44 AA203297	

[1] Numbers 1, 2, 3, etc., represent arbitrary labeling of the 44 nucleotide sequences.

The typical 8-mer has a frequency of about one in 65,000 (4^8) base pairs, or about once in the list of 44 human nuclear receptor nucleotide sequences. The most prevalent 8-mer primer occurred 34 times in the list of

human nuclear receptor sequences shown in Table II. The typical 9-mer has a frequency of about one in 262,000 (4^9). The most prevalent 9-mer primer occurred 20 times in the list of human nuclear receptor sequences shown in Table II.

PCR primers sampling abundant nucleotide sequences were removed from consideration. After removal of abundant nucleotide sequences, 26 of the 100 top ranked 8-mer primers sampling the human nuclear receptor gene family remained in the list of primers. After removal of abundant nucleotide sequences, 50 of the 100 top ranked 9-mer primers sampling the human nuclear receptor gene family remained in the list of primers.

A set of 8-mer primer pairs that amplify all of the human nuclear receptor gene family were identified (see Table III). The simulated PCR products from these 8-mer primer pairs are shown in Table IV.

Table III. 8-mer Primer Pairs for Human Nuclear Receptors

D0' and N0':	TGCAAGGG and TGCAGGAG
D0' and M0':	TGCAAGGG and CAGCAGCG
A0 and R0':	GGCTGCAA and GCCTCCAG
B1' and T0':	TCCTGGAG and CTGCCTGG
E1' and K1 :	CCTTCCTC and CTCCCTGG
M1' and G0 :	CTGCCCTG and AGGGCTGC
W0' and E0 :	CTGCTGGA and CCGCTGCC
C0' and A0':	GGAGGCAG and GGCTGCAA
H1' and O1':	AGCCTGGA and GGGCAGAG
F0 and E1':	GGCAGCTG and GAGGAAGG
F0' and N0':	CAGCTGCC and TGCAGGAG
M3' and G3 :	GATTCCAC and GATGAGCT
D1' and S0 :	CTTCTGGA and CTGGAGCT

Table IV. Simulated PCR Products from 8-mer Primer Pairs for Human Nuclear Receptors

PRIMER PAIR	GENE	SIZE	PRIMERS IN PRODUCT
D0'-N0'	1	702	D0'-N0'
	32	738	D0'-N0'
	12	738	D0'-N0'
	21	846	D0'-N0'
	24	450	D0'-N0'
	36	159	D0'-N0'
	3	843	D0'-N0'
	6	429	D0'-N0'
	29	426	D0'-N0'
D0'-M0'	19	233	D0'-M0'
	20	233	D0'-M0'
	42	812	D0'-M0'
	24	407	D0'-M0'
	36	214	D0'-M0'
	13	254	D0'-M0'
	38	322	D0'-M0'
	18	233	D0'-M0'
	11	254	D0'-M0'
A0-R0'	42	328	A0-R0'
	23	867	A0-R0'
	2	343	A0-R0'
	28	132	A0-R0'
	28	669	R0'-R0'
	30	132	A0-R0'
	30	669	R0'-R0'
	15	803	A0-R0'
B1'-T0'	1	485	B1'-T0'
	36	735	B1'-T0'
	7	114	B1'-T0'
	7	187	T0'-T0'
	13	353	B1'-T0'
	31	350	B1'-T0'
E1'-K1	1	266	E1'-K1
	37	487	E1'-K1
	37	234	K1-K1
	7	228	E1'-K1
	7	398	K1-K1
	26	44	E1'-K1
	27	919	E1'-K1
M1'-G0	1	719	M1'-G0
	34	241	M1'-G0
	8	661	M1'-G0
	17	891	M1'-G0
	43	71	M1'-G0

PRIMER PAIR	GENE	SIZE	PRIMERS IN PRODUCT
W0'-E0	4	589	W0'-E0
	9	950	W0'-E0
	23	144	W0'-E0
	41	144	W0'-E0
	2	706	W0'-E0
C0'-A0'	19	408	C0'-A0'
	20	384	C0'-A0'
	24	257	C0'-A0'
	35	499	C0'-A0'
	14	509	C0'-A0'
H1'-O1'	7	657	H1'-O1'
	2	626	H1'-O1'
	40	126	H1'-O1'
	33	124	H1'-O1'
F0-E1'	1	537	F0-E1'
	9	645	F0-E1'
	16	846	F0-E1'
	10	260	F0-E1'
F0'-N0'	32	327	F0'-N0'
	32	599	F0'-F0'
	35	82	F0'-N0'
	25	978	F0'-N0'
	25	232	N0'-N0'
	22	979	F0'-N0'
M3'-G3	9	224	M3'-G3
	5	891	M3'-G3
	44	498	M3'-G3
D1'-S0	6	907	D1'-S0
	29	904	D1'-S0
	39	361	D1'-S0

The set of 8-mer primers that sampled all (100%) of the identified human nuclear receptor genes contained 21 primers that formed 13 primer pairs. These primers generated 75 simulated PCR products. A set of 9-mer primers that sampled the human nuclear receptor family was also identified and contained 21 primers that formed 12 primer pairs. These 9-mer primers generated 48 simulated PCR products and sample 30 of 44 (68%) of the human nuclear receptor family.

These results demonstrate that a method of the invention can identify a subset of primer pairs that

amplifies all members of the identified group of human nuclear receptor genes.

EXAMPLE III

Selection of PCR Primer Pairs to Amplify Human G-Protein Coupled Receptors

This example demonstrates that a subset of PCR primers that amplifies the human G-protein coupled receptor gene family can be selected using the disclosed method. A computer process was used to generate a set of primer pairs that amplifies the human G-protein coupled receptor gene family (see Example I and Figure 2).

Members of the human G-protein coupled receptor gene family were identified using the human G-protein coupled receptor database internet site (http://receptor.mgh.harvard.edu/species_full.html#026), which is incorporated herein by reference. These genes encode receptors that couple intracellular signaling to GTP-binding proteins. A list of 206 mRNA and mRNA fragments from the human nuclear receptor gene family was compiled. Following removal of duplicates and nucleotide sequences less than 800 bases, 113 human G-protein coupled receptor mRNA sequences remained in the list (see Table V).

The typical 8-mer has a frequency of about one in 65,000 (4^8) base pairs. The most prevalent 8-mer primer occurred 68 times in the list of human G-protein coupled receptor sequences shown in Table V. The typical 9-mer has a frequency of about one in 262,000 (4^9). The most prevalent 9-mer primer occurred 40 times in the list of human G-protein coupled receptor sequences shown in Table V.

Table V. Human G-protein coupled receptor mRNA Sequences and mRNA Fragments[1]

1 D10202	2 D10923	3 D10925	4 D13305	5 D13626
6 D13814	7 D17516	8 D25235	9 D25418	10 D28481
11 D28539	12 D90402	13 J03019	14 J03853	15 L00587
16 L01406	17 L06797	18 L07949	19 L08176	20 L08177
21 L08893	22 L12398	23 L13288	24 L19315	25 L20316
26 L20463	27 L20469	28 L21195	29 L22647	30 L23333
31 L24470	32 L25119	33 L25124	34 L25615	35 L27488
36 L27489	37 L27490	38 L28175	39 M11058	40 M13150
41 M30625	42 M31210	43 M31774	44 M60626	45 M62424
46 M62505	47 M63108	48 M64749	49 M65085	50 M68932
51 M73481	52 M73482	53 M74290	54 M76446	55 M76673
56 M76676	57 M80436	58 M81589	59 M81590	60 M81778
61 M84426	62 M84562	63 M84605	64 M86841	65 M88458
66 M88461	67 M88714	68 M89473	69 M91467	70 M94582
71 S44866	72 S45272	73 S46950	74 S52624	75 S56143
76 S57498	77 S62137	78 S74902	79 S77410	80 U01104
81 U01839	82 U03864	83 U03865	84 U03882	85 U03905
86 U07882	87 U10037	88 U11271	89 U12569	90 U14910
91 U27325	92 X04827	93 X12433	94 X51362	95 X54937
96 X55885	97 X58987	98 X64878	99 X67594	100 X68486
101 X68487	102 X68596	103 X70070	104 X70811	105 X72089
106 X72304	107 X74328	108 X77130	109 X77307	110 X77777
111 X81120	112 X86163	113 Z11687		

[1] Numbers 1, 2, 3, etc., represent arbitrary labeling of the 113 nucleotide sequences.

PCR primers sampling abundant nucleotide sequences were removed from consideration. After removal of abundant nucleotide sequences, 28 of the 100 top ranked 8-mer primers sampling the human G-protein coupled receptor gene family remained in the list of primers. After removal of abundant nucleotide sequences, 45 of the 100 top ranked 9-mer primers sampling the human G-protein

coupled receptor gene family remained in the list of primers

A set of 8-mer primer pairs that amplify all of the human G-protein coupled receptor gene family were identified (see Table VI).

Table VI. 8-mer Primer Pairs for Human G-protein Coupled Receptors

A0' and C0':	GCTGGCCA and TCTGCTGG
M0 and I1 :	ACCTGCTG and CTGCTGGT
L0 and T1 :	TCATCTGC and CTTCCAGG
E0 and L1':	TGCTGGCT and GGCCCTGG
J0' and Z0':	GCTGGTGG and CTCTGCTG
B0 and Z0':	GCTGGCTG and CTCTGCTG
M0 and B1 :	ACCTGCTG and CCAGGGCT
K1' and J0':	GCCCTGGG and CCACCAGC
I0' and D0':	TCCTGCTG and TGGTGGCC
Y0 and F1 :	TCCTGGTG and CTGGGCCA
W0 and A0 :	TGCTGGGC and GCTGGCCA
T0 and F0':	CTGTGCTG and GGCAGCTG
R0' and D1':	CTGGCCAG and CTGGCCAC
A2' and Q2 :	GGCTATGT and TCCAGTCC
G1' and B0 :	CAGCAGCG and CAGCCAGC
T0 and E0':	CTGTGCTG and TGCTGGCT
R1 and C0':	CCTGGCCT and TCTGCTGG
O1' and B0':	TCCTGGCC and GCTGGCTG
B0 and E1 :	GCTGGCTG and TGCCCTGG
L0 and C1':	TCATCTGC and ACAGCAGC
A0' and X0':	GCTGGCCA and GCTGCCTC
O0' and U1':	GCTGCCCC and TGGCCATT
I0' and D0 :	TCCTGCTG and GGCCACCA
G0' and K0 :	CTGCTGCT and CTCCTTCT
F0 and T0':	GGCAGCTG and CTGTGCTG
G0' and S1':	CTGCTGCT and CTGCTGTG
V3' and L0':	GCCTCTGC and TCATCTGC
N1' and B2 :	CTGCCTCC and TGTGGCCC

A0 and T0 : TGGCCAGC and CAGCACAG

The set of 8-mer primers that sampled all (100%) of the identified human G-protein coupled receptor genes contained 45 primers that formed 29 primer pairs. These primers generated 240 simulated PCR products. A set of 9-mer primers that sampled the human G-protein coupled receptor gene family was also identified and contained 53 primers that formed 37 primer pairs. These 9-mer primers generated 178 simulated PCR products and sample 101 of 113 (89%) of the G-protein coupled receptor gene family.

These results demonstrate that a method of the invention can select a subset of primer pairs that amplifies all members of the identified group of human G-protein coupled receptor genes.

EXAMPLE IV

Selection of PCR Primer Pairs to Amplify Human Apoptosis-Associated Genes

This example demonstrates that a subset of PCR primers that amplifies human apoptosis-associated genes can be selected using the disclosed method. A computer process was used to generate a set of primer pairs that amplifies human apoptosis-associated genes (see Example I and Figure 2).

Human apoptosis-associated genes were identified using the PubMed database (http://www4.ncbi.nlm.nih.gov/PubMed/), which is incorporated herein by reference. These genes encode proteins that are known or suspected to be associated with apoptosis, or programmed cell death. A list of 181 mRNA and mRNA fragments of human apoptosis-associated genes was

compiled. Following removal of duplicates and nucleotide sequences less than 800 bases, 60 human apoptosis-associated mRNA sequences remained in the list (see Table VII).

The typical 8-mer has a frequency of about one in 65,000 (4^8) base pairs. The most prevalent 8-mer primer occurred 30 times in the list of human apoptosis-associated gene sequences shown in Table VII. The typical 9-mer has a frequency of about one in 262,000 (4^9). The most prevalent 9-mer primer occurred 17 times in the list of human apoptosis-associated gene sequences shown in Table VII.

Table VII. Human Apoptosis-associated mRNA Sequences and mRNA Fragments[1]

1 D38122	2 D84476	3 L09753	4 Z23115	5 L34583
6 L38509	7 L41690	8 M14745	9 M64722	10 M83554
11 S82185	12 U13022	13 U13670	14 U13696	15 U13697
16 U13737	17 U15172	18 U15173	19 U16811	20 U19251
21 U20536	22 U20537	23 U28015	24 U29656	25 U33201
26 U33286	27 U34584	28 U37449	29 U37518	30 U37546
31 U37547	32 U37688	33 U45878	34 U45880	35 U48869
36 U50062	37 U56976	38 U60519	39 U60520	40 U60521
41 U63295	42 U66879	43 U67319	44 U75381	45 U80017
46 U84388	47 U86214	48 U91985	49 X14454	50 X55313
51 X63717	52 X84709	53 X86779	54 X96710	55 X98172
56 X98176	57 X99404	58 Y07619	59 Y09392	60 Z48810

[1] Numbers 1, 2, 3, etc., represent arbitrary labeling of the 60 nucleotide sequences.

PCR primers sampling abundant nucleotide sequences were removed from consideration. After removal of abundant nucleotide sequences, 15 of the 100 top ranked 8-mer primers sampling the human apoptosis-associated

genes remained in the list of primers. After removal of abundant nucleotide sequences, 30 of the 100 top ranked 9-mer primers sampling the human apoptosis-associated genes remained in the list of primers.

A set of 8-mer primer pairs that amplify 56 of the 60 human apoptosis-associated genes were identified (see Table VIII). Genes 11, 13, 25 and 54 from the list in Table VII were not amplified.

The set of 8-mer primers that sampled 56 of 60 (93%) of the identified human apoptosis-associated genes contained 42 primers that formed 24 primer pairs. These primers generated 106 simulated PCR products. A set of 9-mer primers that sampled the human apoptosis-associated

Table VIII. 8-mer Primer Pairs for Human Apoptosis-associated Genes

K1'and G0 :	CTGGAGGA	and TCATCCAG
F0 and A0 :	CTGGAGAA	and GCTGCAGC
E0 and P2 :	CTGCTGGA	and GAACAGGA
J0' and V0':	GCTCCTGG	and GCCCCTGG
Z0' and K0':	CCAGAGCA	and CAGCCCAG
J2 and U1':	CTTCCTGA	and TGGATGCA
L1 and F0':	TCCAGTTC	and CTGGAGAA
D0 and E0':	TGAAGAGC	and CTGCTGGA
A0' and A0':	CCTGGGAG	and CTCCCAGG
L0' and W0 :	CCAGGCAG	and CAGAGGAG
B0' and X0 :	CCTCCAGG	and GGAGGCAG
H2 and J1 :	GGCGGCTG	and TCTGCTGG
O2' and P2 :	GCTCGGCC	and GAACAGGA
F0 and Z0 :	CTGGAGAA	and CCAGAGCA
O0 and H0':	GCTGGTGG	and GGCAGCTG
W0' and O3 :	CAGAGGAG	and TCCTGGGT
B0' and X0':	CCTCCAGG	and CTGCCTCC
B1 and L1 :	CTCAGCAG	and GAACTGGA
T1' and H2 :	CAGCTGGA	and CAGCCGCC
U2 and J3 :	CTGCATCC	and TGCAACAG

H3 and V0': GGCTGCGC and GCCCCTGG
P1' and Y2': CTGGCCAG and GTGGTGGA
R1' and G1 : CCTGCAGG and GCCTCCAG

genes was also identified and contained 24 primers that formed 12 primer pairs. These 9-mer primers generated 41 simulated PCR products and sampled 29 of 60 (48%) of the human apoptosis-associated genes.

These results demonstrate that a method of the invention can select a subset of primer pairs that amplifies 56 of the 60 members of the identified group of human apoptosis-associated genes.

EXAMPLE V

Selection of PCR Primer Pairs to Amplify Human DNA Repair and Replication Genes

This example demonstrates that a subset of PCR primers that amplifies human DNA repair and replication genes can be selected using the disclosed method. A computer process was used to generate a set of primer pairs that amplifies human DNA repair and replication genes (see Example I and Figure 2).

Human DNA repair and replication genes were identified using the PubMed database (http://www4.ncbi.nlm.nih.gov/PubMed/), which is incorporated herein by reference. A list of 169 mRNA and mRNA fragments of human DNA repair and replication genes was compiled. Following removal of duplicates and nucleotide sequences less than 800 bases, 65 human DNA repair and replication gene mRNA sequences remained in the list (see Table IX).

The typical 8-mer has a frequency of about one in 65,000 (4^8) base pairs. The most prevalent 8-mer primer occurred 36 times in the list of human DNA repair and replication gene sequences shown in Table IX. The typical 9-mer has a frequency of about one in 262,000 (4^9). The most prevalent 9-mer primer occurred 22 times in the list of human DNA repair and replication gene sequences shown in Table IX.

Table IX. Human DNA Repair and Replication Genes mRNA Sequences and mRNA Fragments[1]

1 D64108	2 X83441	3 L24444	4 X91992	5 D13370
6 Z30094	7 L27425	8 X84740	9 M28650	10 X52221
11 M31899	12 L04791	13 D21235	14 D21090	15 M74524
16 M74525	17 M36067	18 X69821	19 M29971	20 L47579
21 X81030	22 D38500	23 D38501	24 D38502	25 D38503
26 D13804	27 L33262	28 X97795	29 M29474	30 M94633
31 L07872	32 X78627	33 X78262	34 U61981	35 U63139
36 U63329	37 U64315	38 D14533	39 Z11495	40 D21089
41 X69978	42 X71342	43 L34079	44 K03199	45 L09561
46 L37374	47 M87499	48 U09559	49 U12134	50 U13695
51 U13696	52 U18300	53 U27346	54 U28946	55 U32986
56 U37359	57 U40622	58 U40671	59 U47077	60 U72936
61 U75967	62 X15653	63 X83753	64 Y10658	65 Z48796

[1] Numbers 1, 2, 3, etc., represent arbitrary labeling of the 65 nucleotide sequences.

PCR primers sampling abundant nucleotide sequences were removed from consideration. After removal of abundant nucleotide sequences, 15 of the 100 top ranked 8-mer primers sampling the human DNA repair and replication genes remained in the list of primers. After removal of abundant nucleotide sequences, 38 of the 100 top ranked 9-mer primers sampling the human DNA repair and replication genes remained in the list of primers.

A set of 8-mer primer pairs that amplify 64 of the 65 human DNA repair and replication genes were identified (see Table X). Gene 38 from the list in Table IX was not amplified.

Table X. 8-mer Primer Pairs for Human DNA Repair and Replication Genes

E0' and A1':	GGAAGGAG	and TGCAGGAG
H0 and K1 :	CTGGCTGA	and CTTCCTCA
N0' and V0':	TCATCCAG	and AGCAGCAA
F0 and E1 :	AGGCTGGG	and AGAAGGAG
W1' and Q0':	CTGGAGGA	and AGCTGAAG
D0 and A0 :	CATCCAGA	and GCTGCAGC
T0 and B0':	CTTCCTGA	and CCTCCTGG
L1' and X1 :	TGCTCTGG	and CTGCTGAA
H0 and Z0 :	CTGGCTGA	and GCTGCTGA
R0 and A1':	TCCAGAAG	and TGCAGGAG
K0' and C0':	GCAGCTGC	and CTGCTGGA
A0' and I1':	GCTGCAGC	and GGCAGAGG
O3 and D3 :	AGTCTGGA	and AGTTTGCC
E1' and T1 :	AGAAGGAG	and GTCCAGGT
C1 and N2':	CACAGCTG	and TTGCTGAC
I2 and C2':	CTCAGTGC	and TGCCACCA
D2 and Z2':	AGAACCTG	and ACCATTGC
E1' and Z0 :	AGAAGGAG	and GCTGCTGA
H1 and B0 :	GGAGGCAG	and CCAGGAGG
G2' and O2':	TGGAGAGA	and CTGATGAC
U1 and Q1':	GAGATGGA	and AGATGCTG
A0' and B1 :	GCTGCAGC	and GCTGGAAG
G0' and S1 :	GCCAGGAG	and CAGGAAGC
B0 and O0 :	CCTCCTGG	and TCTGGAAG
N3 and K1':	CATCTGGA	and TGAGGAAG

The set of 8-mer primers that sampled 64 of 65 (98%) of the identified human DNA repair and replication genes contained 44 primers that formed 25 primer pairs. These primers generated 120 simulated PCR products. A set of 9-mer primers that sampled the human DNA repair and

replication genes was also identified and contained 28 primers that formed 15 primer pairs. These 9-mer primers generated 51 simulated PCR products and sampled 35 of 65 (54%) of the human DNA repair and replication genes.

These results demonstrate that a method of the invention can select a subset of primer pairs that amplifies 64 of the 65 members of the identified group of human DNA repair and replication genes.

Although the invention has been described with reference to the disclosed embodiments, those skilled in the art will readily appreciate that the specific experiments detailed are only illustrative of the invention. It should be understood that various modifications can be made without departing from the spirit of the invention. Accordingly, the invention is limited only by the following claims.

We claim:

1. A subset of primers sufficient to amplify a group of at least 10 related nucleotide sequences.

2. The subset of primers of claim 1, wherein the number of primers in said subset is less than or equal to the number of related nucleotide sequences in said group.

3. The subset of primers of claim 2, wherein said primers are about 5 to about 50 nucleotides in length.

4. The subset of primers of claim 3, wherein said primers are about 8 to about 12 nucleotides in length.

5. The subset of primers of claim 2, wherein said primers have a predetermined G+C content.

6. The subset of primers of claim 5, wherein said primers have G+C content of about 50% to about 90%.

7. The subset of primers of claim 2, wherein said subset of primers is a near minimal subset of primers.

8. The subset of primers of claim 2, wherein said subset of primers is a minimal subset of primers.

9. The subset of primers of claim 2, wherein said subset of primers is sufficient to amplify all of said related nucleotide sequences.

10. The subset of primers of claim 2, wherein said subset of primers excludes primer pairs that amplify an undesirable nucleotide sequence.

11. The subset of primers of claim 10, wherein said subset of primers excludes primers that match an undesirable nucleotide sequence.

12. The subset of primers of claim 10, wherein said undesirable nucleotide sequence is selected from the group consisting of ribosomal RNA, mitochondrial DNA and a dispersed repetitive sequence element.

13. The subset of primers of claim 2, wherein said group of related nucleotide sequences is selected from the group consisting of nuclear receptor genes, G-protein coupled receptor genes, apoptosis-associated genes, and DNA repair and replication genes.

14. The subset of primers of claim 13, wherein said group of related nucleotide sequences comprises nuclear hormone receptor genes.

15. The subset of primers of claim 14, wherein said subset comprises a primer selected from the group consisting of TGCAAGGG; TGCAGGAG; CAGCAGCG; GGCTGCAA; GCCTCCAG; TCCTGGAG; CTGCCTGG; CCTTCCTC; CTCCCTGG; CTGCCCTG; AGGGCTGC; CTGCTGGA; CCGCTGCC; GGAGGCAG; AGCCTGGA; GGGCAGAG; GGCAGCTG; GAGGAAGG; CAGCTGCC; GATTCCAC; GATGAGCT; CTTCTGGA; and CTGGAGCT.

16. The subset of primers of claim 14, wherein said subset comprises the primers TGCAAGGG; TGCAGGAG; CAGCAGCG; GGCTGCAA; GCCTCCAG; TCCTGGAG; CTGCCTGG; CCTTCCTC; CTCCCTGG; CTGCCCTG; AGGGCTGC; CTGCTGGA; CCGCTGCC; GGAGGCAG; AGCCTGGA; GGGCAGAG; GGCAGCTG;

GAGGAAGG; CAGCTGCC; GATTCCAC; GATGAGCT; CTTCTGGA; and CTGGAGCT.

17. The subset of primers of claim 16, wherein said subset comprises the primer pairs TGCAAGGG and TGCAGGAG; TGCAAGGG and CAGCAGCG; GGCTGCAA and GCCTCCAG; TCCTGGAG and CTGCCTGG; CCTTCCTC and CTCCCTGG; CTGCCCTG and AGGGCTGC; CTGCTGGA and CCGCTGCC; GGAGGCAG and GGCTGCAA; AGCCTGGA and GGGCAGAG; GGCAGCTG and GAGGAAGG; CAGCTGCC and TGCAGGAG; GATTCCAC and GATGAGCT; and CTTCTGGA and CTGGAGCT.

18. The subset of primers of claim 13, wherein said group of related nucleotide sequences comprises G-protein coupled receptors.

19. The subset of primers of claim 18, wherein said subset comprises a primer selected from the group consisting of GCTGGCCA; TCTGCTGG; CTGTGCTG; TGCTGGCT; ACCTGCTG; CTGCTGGT; CCTGGCCT; TCATCTGC; CTTCCAGG; TCCTGGCC; GCTGGCTG; GGCCCTGG; TGCCCTGG; GCTGGTGG; CTCTGCTG; ACAGCAGC; GCTGCCTC; CCAGGGCT; GCTGCCCC; TGGCCATT; GCCCTGGG; CCACCAGC; TCCTGCTG; GGCCACCA; TGGTGGCC; CTGCTGCT; CTCCTTCT; TCCTGGTG; CTGGGCCA; GGCAGCTG; TGCTGGGC; CTGCTGTG; GCCTCTGC; CTGGCCAG; CTGGCCAC; CTGCCTCC; TGTGGCCC; GGCTATGT; TCCAGTCC; TGGCCAGC; CAGCACAG; CAGCAGCG; and CAGCCAGC.

20. The subset of primers of claim 18, wherein said subset comprises the primers GCTGGCCA; TCTGCTGG; CTGTGCTG; TGCTGGCT; ACCTGCTG; CTGCTGGT; CCTGGCCT; TCATCTGC; CTTCCAGG; TCCTGGCC; GCTGGCTG; GGCCCTGG; TGCCCTGG; GCTGGTGG; CTCTGCTG; ACAGCAGC; GCTGCCTC; CCAGGGCT; GCTGCCCC; TGGCCATT; GCCCTGGG; CCACCAGC;

TCCTGCTG; GGCCACCA; TGGTGGCC; CTGCTGCT; CTCCTTCT; TCCTGGTG; CTGGGCCA; GGCAGCTG; TGCTGGGC; CTGCTGTG; GCCTCTGC; CTGGCCAG; CTGGCCAC; CTGCCTCC; TGTGGCCC; GGCTATGT; TCCAGTCC; TGGCCAGC; CAGCACAG; CAGCAGCG; and CAGCCAGC.

21. The subset of primers of claim 20, wherein said subset comprises the primer pairs GCTGGCCA and TCTGCTGG; CTGTGCTG and TGCTGGCT; ACCTGCTG and CTGCTGGT; CCTGGCCT and TCTGCTGG; TCATCTGC and CTTCCAGG; TCCTGGCC and GCTGGCTG; TGCTGGCT and GGCCCTGG; GCTGGCTG and TGCCCTGG; GCTGGTGG and CTCTGCTG; TCATCTGC and ACAGCAGC; GCTGGCTG and CTCTGCTG; GCTGGCCA and GCTGCCTC; ACCTGCTG and CCAGGGCT; GCTGCCCC and TGGCCATT; GCCCTGGG and CCACCAGC; TCCTGCTG and GGCCACCA; TCCTGCTG and TGGTGGCC; CTGCTGCT and CTCCTTCT; TCCTGGTG and CTGGGCCA; GGCAGCTG and CTGTGCTG; TGCTGGGC and GCTGGCCA; CTGCTGCT and CTGCTGTG; CTGTGCTG and GGCAGCTG; GCCTCTGC and TCATCTGC; CTGGCCAG and CTGGCCAC; CTGCCTCC and TGTGGCCC; GGCTATGT and TCCAGTCC; TGGCCAGC and CAGCACAG; and CAGCAGCG and CAGCCAGC.

22. The subset of primers of claim 13, wherein said group of related nucleotide sequences comprises apoptosis-associated genes.

23. The subset of primers of claim 22, wherein said subset comprises a primer selected from the group consisting of CTGGAGGA; TCATCCAG; CTGGAGAA; GCTGCAGC; CTGCTGGA; GAACAGGA; GCTCCTGG; GCCCCTGG; CCAGAGCA; CAGCCCAG; CTTCCTGA; TGGATGCA; TCCAGTTC; TGAAGAGC; CCTGGGAG; CTCCCAGG; CCAGGCAG; CAGAGGAG; CCTCCAGG; GGAGGCAG; GGCGGCTG; TCTGCTGG; GCTCGGCC; GCTGGTGG;

GGCAGCTG; TCCTGGGT; CTGCCTCC; CTCAGCAG; GAACTGGA; CAGCTGGA; CAGCCGCC; CTGCATCC; TGCAACAG; GGCTGCGC; CTGGCCAG; GTGGTGGA; CCTGCAGG; and GCCTCCAG.

24. The subset of primers of claim 22, wherein said subset comprises the primers CTGGAGGA; TCATCCAG; CTGGAGAA; GCTGCAGC; CTGCTGGA; GAACAGGA; GCTCCTGG; GCCCCTGG; CCAGAGCA; CAGCCCAG; CTTCCTGA; TGGATGCA; TCCAGTTC; TGAAGAGC; CCTGGGAG; CTCCCAGG; CCAGGCAG; CAGAGGAG; CCTCCAGG; GGAGGCAG; GGCGGCTG; TCTGCTGG; GCTCGGCC; GCTGGTGG; GGCAGCTG; TCCTGGGT; CTGCCTCC; CTCAGCAG; GAACTGGA; CAGCTGGA; CAGCCGCC; CTGCATCC; TGCAACAG; GGCTGCGC; CTGGCCAG; GTGGTGGA; CCTGCAGG; and GCCTCCAG.

25. The subset of primers of claim 24, wherein said subset comprises the primer pairs CTGGAGGA and TCATCCAG; CTGGAGAA and GCTGCAGC; CTGCTGGA and GAACAGGA; GCTCCTGG and GCCCCTGG; CCAGAGCA and CAGCCCAG; CTTCCTGA and TGGATGCA; TCCAGTTC and CTGGAGAA; TGAAGAGC and CTGCTGGA; CCTGGGAG and CTCCCAGG; CCAGGCAG and CAGAGGAG; CCTCCAGG and GGAGGCAG; GGCGGCTG and TCTGCTGG; GCTCGGCC and GAACAGGA; CTGGAGAA and CCAGAGCA; GCTGGTGG and GGCAGCTG; CAGAGGAG and TCCTGGGT; CCTCCAGG and CTGCCTCC; CTCAGCAG and GAACTGGA; CAGCTGGA and CAGCCGCC; CTGCATCC and TGCAACAG; GGCTGCGC and GCCCCTGG; CTGGCCAG and GTGGTGGA; and CCTGCAGG and GCCTCCAG.

26. The subset of primers of claim 13, wherein said group of related nucleotide sequences comprises DNA repair and replication genes.

27. The subset of primers of claim 26, wherein said subset comprises a primer selected from the group consisting of GGAAGGAG; TGCAGGAG; CTGGCTGA; CTTCCTCA; TCATCCAG; AGCAGCAA; AGGCTGGG; AGAAGGAG; CTGGAGGA; AGCTGAAG; CATCCAGA; GCTGCAGC; CTTCCTGA; CCTCCTGG; TGCTCTGG; CTGCTGAA; GCTGCTGA; TCCAGAAG; GCAGCTGC; CTGCTGGA; GGCAGAGG; AGTCTGGA; AGTTTGCC; GTCCAGGT; CACAGCTG; TTGCTGAC; CTCAGTGC; TGCCACCA; AGAACCTG; ACCATTGC; GGAGGCAG; CCAGGAGG; TGGAGAGA; CTGATGAC; GAGATGGA; AGATGCTG; GCTGGAAG; GCCAGGAG; CAGGAAGC; TCTGGAAG; CATCTGGA; and TGAGGAAG.

28. The subset of primers of claim 26, wherein said subset comprises the primers GGAAGGAG; TGCAGGAG; CTGGCTGA; CTTCCTCA; TCATCCAG; AGCAGCAA; AGGCTGGG; AGAAGGAG; CTGGAGGA; AGCTGAAG; CATCCAGA; GCTGCAGC; CTTCCTGA; CCTCCTGG; TGCTCTGG; CTGCTGAA; GCTGCTGA; TCCAGAAG; GCAGCTGC; CTGCTGGA; GGCAGAGG; AGTCTGGA; AGTTTGCC; GTCCAGGT; CACAGCTG; TTGCTGAC; CTCAGTGC; TGCCACCA; AGAACCTG; ACCATTGC; GGAGGCAG; CCAGGAGG; TGGAGAGA; CTGATGAC; GAGATGGA; AGATGCTG; GCTGGAAG; GCCAGGAG; CAGGAAGC; TCTGGAAG; CATCTGGA; and TGAGGAAG.

29. The subset of primers of claim 28, wherein said subset comprises the primer pairs GGAAGGAG and TGCAGGAG; CTGGCTGA and CTTCCTCA; TCATCCAG and AGCAGCAA; AGGCTGGG and AGAAGGAG; CTGGAGGA and AGCTGAAG; CATCCAGA and GCTGCAGC; CTTCCTGA and CCTCCTGG; TGCTCTGG and CTGCTGAA; CTGGCTGA and GCTGCTGA; TCCAGAAG and TGCAGGAG; GCAGCTGC and CTGCTGGA; GCTGCAGC and GGCAGAGG; AGTCTGGA and AGTTTGCC; AGAAGGAG and GTCCAGGT; CACAGCTG and TTGCTGAC; CTCAGTGC and TGCCACCA; AGAACCTG and ACCATTGC; AGAAGGAG and GCTGCTGA; GGAGGCAG and CCAGGAGG; TGGAGAGA and CTGATGAC; GAGATGGA and AGATGCTG;

GCTGCAGC and GCTGGAAG; GCCAGGAG and CAGGAAGC; CCTCCTGG and TCTGGAAG; and CATCTGGA and TGAGGAAG.

30. A subset of primers sufficient to amplify a group of related nucleotide sequences,

wherein said subset comprises at least one anchor primer of the formula T_nX_m, wherein X is selected from the group consisting of G, A, C and T, n is a number between 10 and 20 and m is a number between 0 and 3; and

wherein said subset comprises one or more second primers, wherein said second primer combined with said anchor primer amplifies two or more related nucleotide sequences in said group.

31. The subset of primers of claim 30, wherein said group comprises at least 10 related nucleotide sequences.

32. A method of selecting a subset of primers sufficient to amplify a group of related nucleotide sequences, comprising the steps of:

a) identifying the group of related nucleotide sequences;

b) generating a set of primers that match each of the related nucleotide sequences in said group;

c) determining for each systematic pairing of each primer, which of said related nucleotide sequences are amplified; and

d) selecting from said systematic pairing, a subset of primers, wherein said selected subset of primers is sufficient to amplify said group of related nucleotide sequences.

33. The method according to claim 32, wherein said primers are about 5 to about 50 nucleotides in length.

34. The method according to claim 33, wherein said primers are about 8 to about 12 nucleotides in length.

35. The method according to claim 32, wherein said primers have a predetermined G+C content.

36. The method according to claim 35, wherein said primers have G+C content of about 50% to about 90%.

37. The method according to claim 32, wherein said subset of primers is a near minimal subset of primers.

38. The method according to claim 32, wherein said subset of primers is a minimal subset of primers.

39. The method according to claim 32, wherein said subset of primers contains all primers sufficient to amplify said group of related nucleotide sequences.

40. The method according to claim 32, wherein said group of related nucleotide sequences is a group of at least 10 related nucleotide sequences.

41. The method according to claim 32, wherein the related nucleotide sequences in said group are structurally related.

42. The method according to claim 41, wherein said group of related nucleotide sequences is selected from the group consisting of nuclear receptor genes and G-protein coupled receptor genes.

43. The method according to claim 32, wherein said group of related nucleotide sequences is selected from the group consisting of apoptosis-associated genes and DNA repair and replication genes.

44. The method according to claim 32, wherein said selected subset of primers is sufficient to amplify all of said related nucleotide sequences.

45. The method according to claim 32, wherein said subset of primers excludes primer pairs that amplify an undesirable nucleotide sequence.

46. The method according to claim 45, wherein said undesirable nucleotide sequence is selected from the group consisting of ribosomal RNA, mitochondrial DNA and a dispersed repetitive sequence element.

47. The method according to claim 32, wherein said subset of primers excludes primers that match an undesirable nucleotide sequence.

48. The method according to claim 47, wherein said undesirable nucleotide sequence is selected from the group consisting of ribosomal RNA, mitochondrial DNA and a dispersed repetitive sequence element.

49. A subset of primers sufficient to amplify a group of related nucleotide sequences, said subset of primers produced by the method of claim 32.

50. The subset of primers of claim 49, wherein said group of related nucleotide sequences are selected from the group consisting of nuclear receptor genes, G-protein coupled receptor genes, apoptosis-associated genes, and DNA repair and replication genes.

51. A method of identifying an amplified nucleotide sequence related to an original group of related nucleotide sequences, comprising the steps of:

a) amplifying a population of nucleotide sequences with the subset of primers of claim 47; and

b) identifying amplified nucleotide sequences, wherein said amplified nucleotide sequences are related to the original group of related nucleotide sequences.

52. The method of claim 51, wherein at least one of said amplified nucleotide sequences was not in the original group of related nucleotide sequences.

53. In a computer, a method of selecting a subset of primers sufficient to amplify a group of related nucleotide sequences, comprising the computer-executed steps of:

a) identifying the group of related nucleotide sequences;

b) generating a set of primers that match each of the related nucleotide sequences in said group;

c) determining for each systematic pairing of each primer, which of said related nucleotide sequences are amplified; and

d) selecting from said systematic pairing, a subset of primers, wherein said selected subset of primers is sufficient to amplify said group of related nucleotide sequences.

54. The method according to claim 53, wherein said primers are about 5 to about 50 nucleotides in length.

55. The method according to claim 53, wherein said primers are about 8 to about 12 nucleotides in length.

56. The method according to claim 53, wherein said primers have a predetermined G+C content.

57. The method according to claim 56, wherein said primers have G+C content of about 50% to about 90%.

58. The method according to claim 53, wherein said subset of primers is a near minimal subset of primers.

59. The method according to claim 53, wherein said subset of primers is a minimal subset of primers.

60. The method according to claim 53, wherein said subset of primers contains all primers sufficient to amplify said group of related nucleotide sequences.

61. The method according to claim 53, wherein said group of related nucleotide sequences is a group of at least 10 nucleotide sequences.

62. The method according to claim 53, wherein the related nucleotide sequences in said group are structurally related.

63. The method according to claim 62, wherein the group of related nucleotide sequences is selected from the group consisting of nuclear receptor genes and G-protein coupled receptor genes.

64. The method according to claim 53, wherein said group of related nucleotide sequences is selected from the group consisting of apoptosis-associated genes, and DNA repair and replication genes.

65. The method according to claim 53, wherein said selected subset of primers is sufficient to amplify all of said related nucleotide sequences.

66. The method according to claim 53, wherein said subset of primers excludes primer pairs that amplify an undesirable nucleotide sequence.

67. The method according to claim 66, wherein said undesirable nucleotide sequence is selected from the group consisting of ribosomal RNA, mitochondrial DNA and a dispersed repetitive sequence element.

68. The method according to claim 53, wherein said subset of primers excludes primers that match an undesirable nucleotide sequence.

69. The method according to claim 68, wherein said undesirable nucleotide sequence is selected from the group consisting of ribosomal RNA, mitochondrial DNA and a dispersed repetitive sequence element.

70. A subset of primers sufficient to amplify a group of related nucleotide sequences, said subset of primers produced by the method of claim 53.

71. The subset of primers of claim 70, wherein said group of related nucleotide sequences is selected from the group consisting of nuclear receptor genes, G-protein coupled receptor genes, apoptosis-associated genes, and DNA repair and replication genes.

72. A computer apparatus, comprising:
a processor;

main memory in communication with said processor;

a primer pair selector in communication with said main memory for carrying out the computer-executed steps of:

a) identifying a group of related nucleotide sequences;

b) generating a set of primers that matches each of the related nucleotide sequences in said group;

c) determining for each systematic pairing of each primer, which of said related nucleotide sequences are amplified; and

d) selecting from said systematic pairing, a subset of primers, wherein said selected subset of primers is sufficient to amplify the group of related nucleotide sequences.

73. A computer program product for determining a set of primers sufficient to amplify a group of related nucleotide sequences, comprising:

a) means for identifying a group of related nucleotide sequences;

b) means for generating a set of primers that matches each of said related nucleotide sequences;

c) means for determining for each systematic pairing of each primer, which of said related nucleotide sequences are amplified;

d) means for selecting from said systematic pairings a subset of primers which can amplify the related nucleotide sequences; and

e) signal-bearing media containing said means for the identifying, generating, determining and selecting.

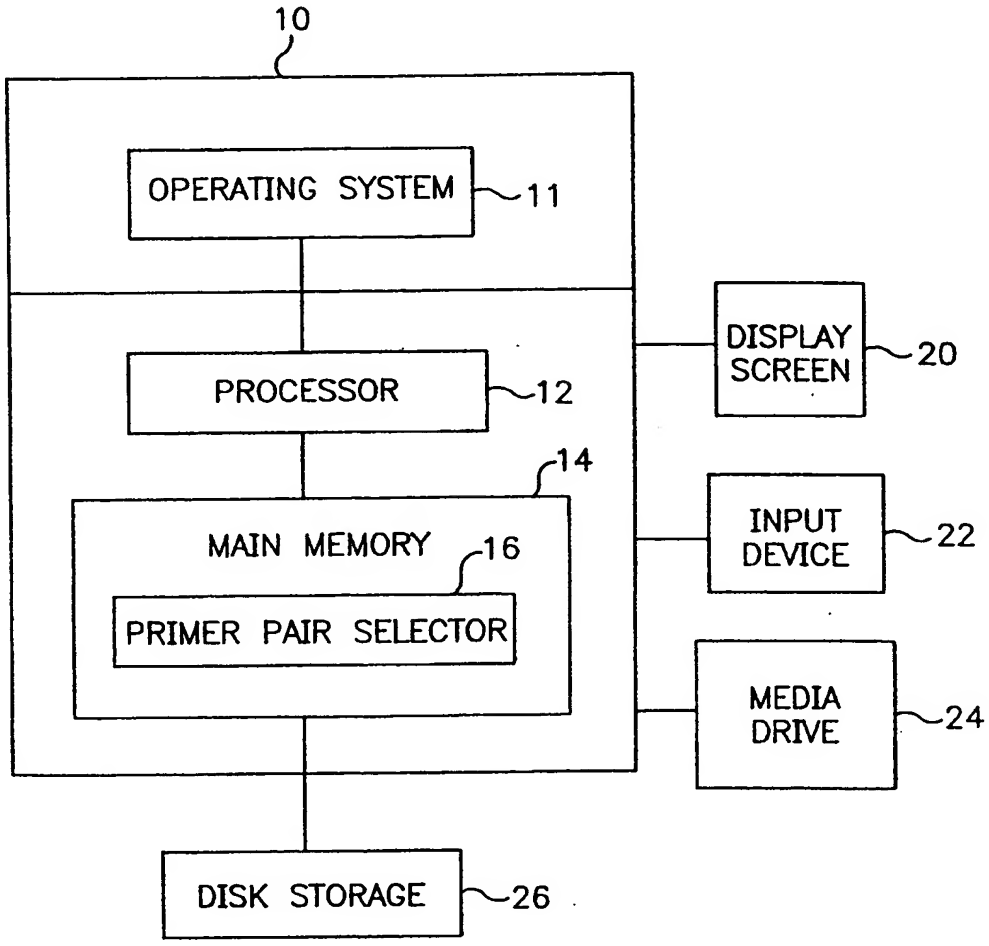
10
OPERATING SYSTEM
11
PROCESSOR
12
14
MAIN MEMORY
16
PRIMER PAIR SELECTOR
DISK STORAGE
26
DISPLAY SCREEN
20
INPUT DEVICE
22
MEDIA DRIVE
24

FIG. 1

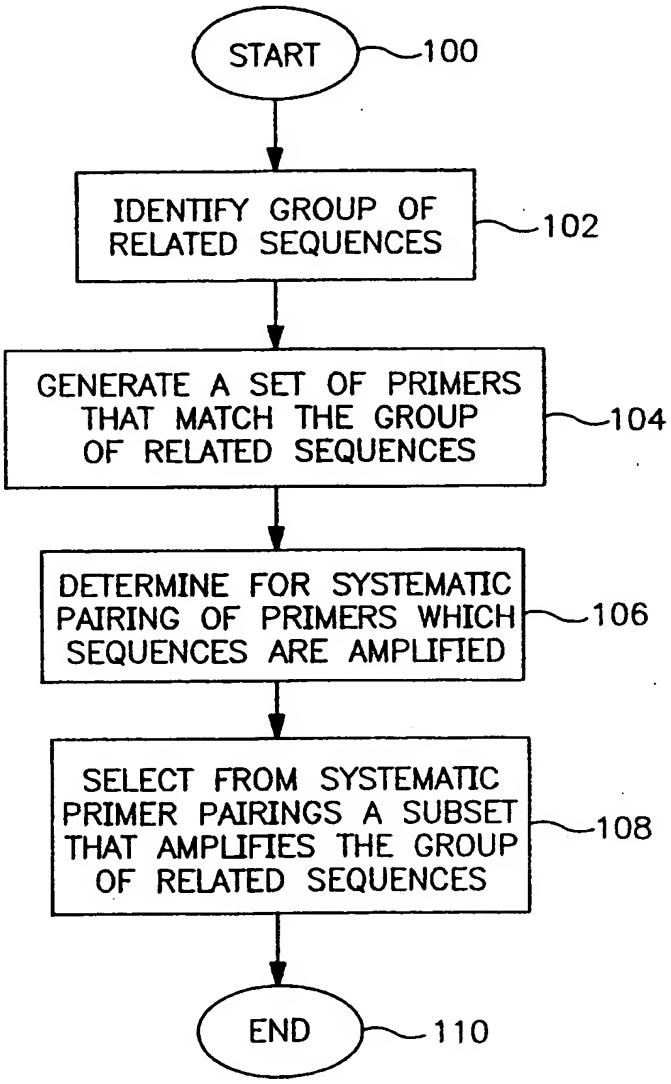
START
100
IDENTIFY GROUP OF RELATED SEQUENCES
102
GENERATE A SET OF PRIMERS THAT MATCH THE GROUP OF RELATED SEQUENCES
104
DETERMINE FOR SYSTEMATIC PAIRING OF PRIMERS WHICH SEQUENCES ARE AMPLIFIED
106
SELECT FROM SYSTEMATIC PRIMER PAIRINGS A SUBSET THAT AMPLIFIES THE GROUP OF RELATED SEQUENCES
108
END
110

FIG. 2

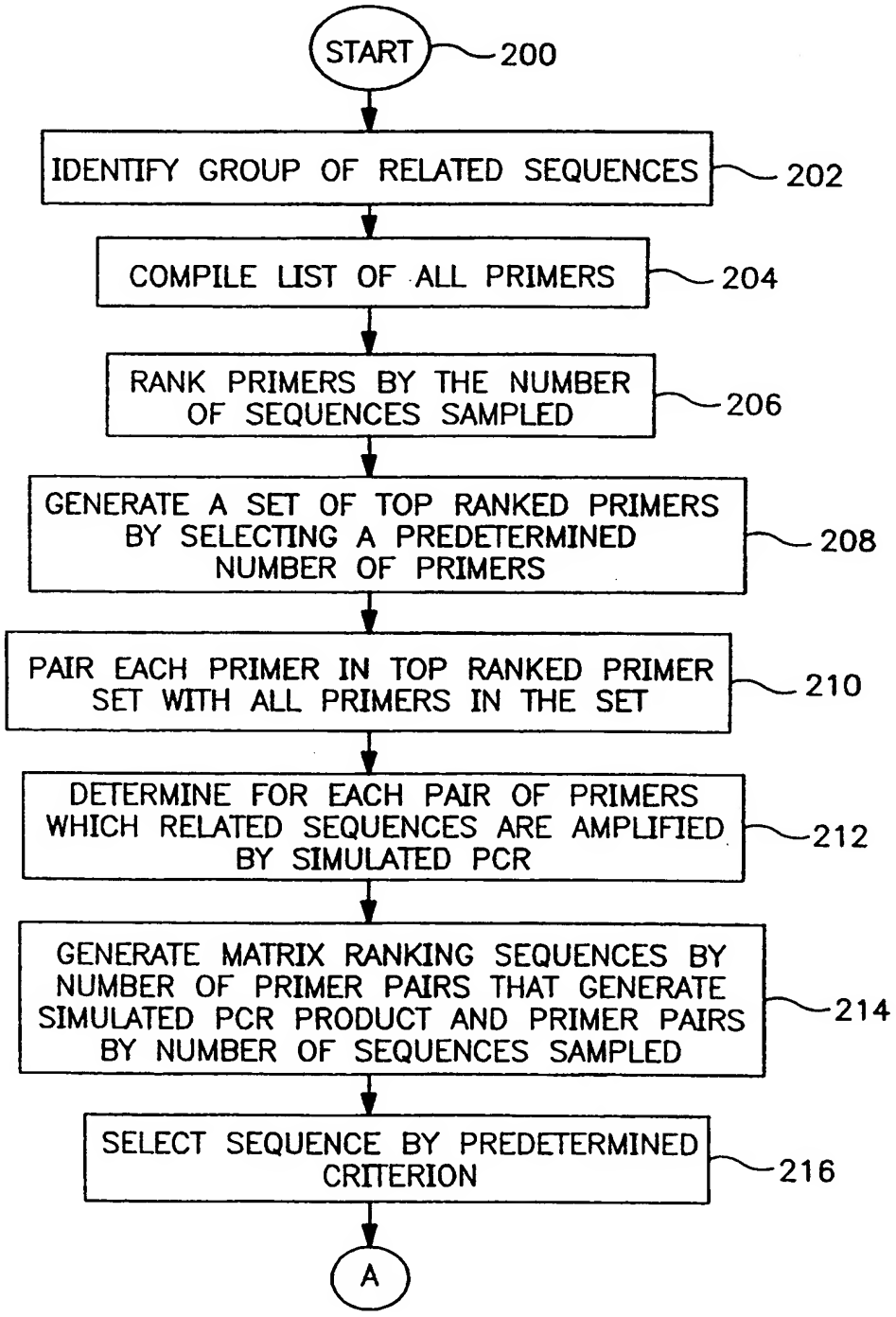
START
200
IDENTIFY GROUP OF RELATED SEQUENCES
202
COMPILE LIST OF ALL PRIMERS
204
RANK PRIMERS BY THE NUMBER OF SEQUENCES SAMPLED
206
GENERATE A SET OF TOP RANKED PRIMERS BY SELECTING A PREDETERMINED NUMBER OF PRIMERS
208
PAIR EACH PRIMER IN TOP RANKED PRIMER SET WITH ALL PRIMERS IN THE SET
210
DETERMINE FOR EACH PAIR OF PRIMERS WHICH RELATED SEQUENCES ARE AMPLIFIED BY SIMULATED PCR
212
GENERATE MATRIX RANKING SEQUENCES BY NUMBER OF PRIMER PAIRS THAT GENERATE SIMULATED PCR PRODUCT AND PRIMER PAIRS BY NUMBER OF SEQUENCES SAMPLED
214
SELECT SEQUENCE BY PREDETERMINED CRITERION
216
A

FIG. 3

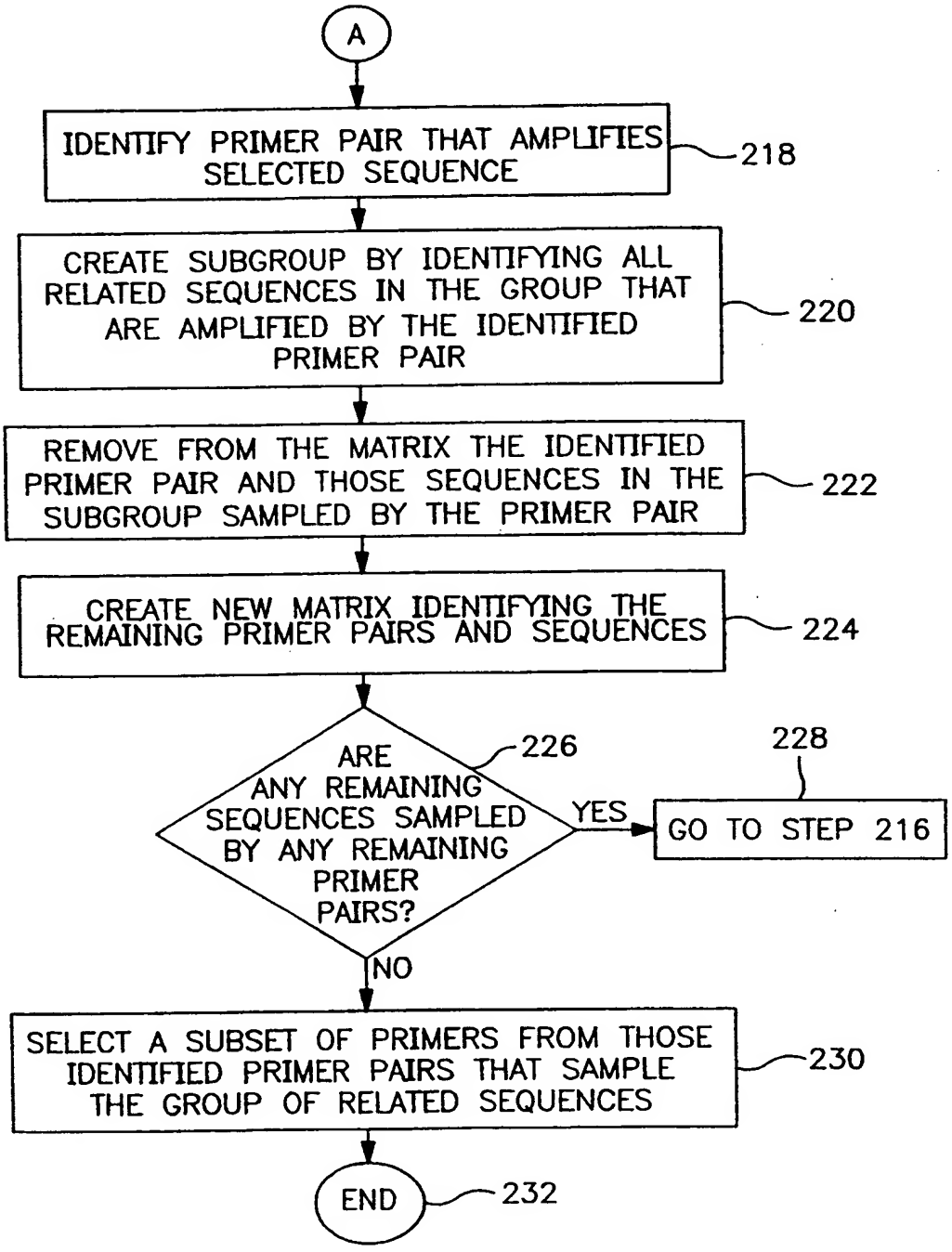
A
IDENTIFY PRIMER PAIR THAT AMPLIFIES SELECTED SEQUENCE
218
CREATE SUBGROUP BY IDENTIFYING ALL RELATED SEQUENCES IN THE GROUP THAT ARE AMPLIFIED BY THE IDENTIFIED PRIMER PAIR
220
REMOVE FROM THE MATRIX THE IDENTIFIED PRIMER PAIR AND THOSE SEQUENCES IN THE SUBGROUP SAMPLED BY THE PRIMER PAIR
222
CREATE NEW MATRIX IDENTIFYING THE REMAINING PRIMER PAIRS AND SEQUENCES
224
ARE ANY REMAINING SEQUENCES SAMPLED BY ANY REMAINING PRIMER PAIRS?
226
YES
228
GO TO STEP 216
NO
SELECT A SUBSET OF PRIMERS FROM THOSE IDENTIFIED PRIMER PAIRS THAT SAMPLE THE GROUP OF RELATED SEQUENCES
230
END
232

FIG. 4

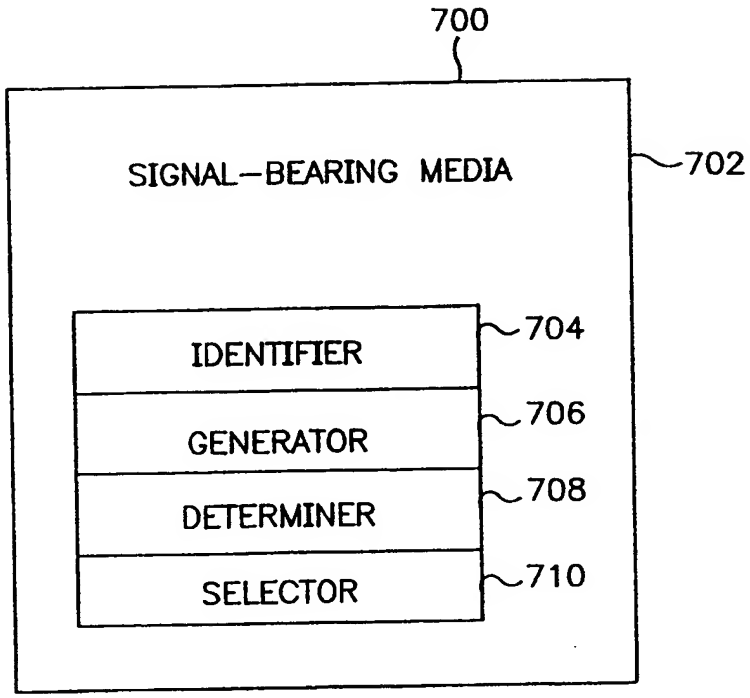
700
SIGNAL-BEARING MEDIA
702
IDENTIFIER
704
GENERATOR
706
DETERMINER
708
SELECTOR
710

FIG.5